electronics

MDPI

Article

Refusal, Confidence, and Explanation: A Cross-Modal Audit of Uncertainty-Expression Failures in Small Open-Weight Language and Vision-Language Models

Minjong Cheon

Department of Computer Science, Korea National Open University, Seoul 03087, Republic of Korea; jmj2316@mail.knou.ac.kr

Abstract

Safety evaluations usually test refusal, abstention, and stated confidence separately. We treat them as one property, appropriate uncertainty expression, and audit it in text and vision across six small open-weight LLMs and five vision-language model (VLM) configurations. On a 240-item Korean benchmark, judged refusal is higher for harmful than for benign-looking requests in every model and under every labeling rule, but model-level orderings depend on the rule (primary rule: zephyr-7b-beta refuses harmful requests least, 42.5%, and benign ones most, 18.8%). Against 150 blinded labels from one annotator, the judges agree on refusal ($\kappa = 0.63$) but overstate benign-prompt refusal (10.4% vs. 3.6%). Replacing sensitive words in benign prompts did not lower refusal (-1.9 pp $[-5.4, 1.7]$). Two VLMs abstain far more on out-of-distribution images, yet their stated confidence when they answer is equivalent within ± 5 points to that on in-distribution images. On six MedMNIST v2 tasks, VLM accuracy is near chance (13.2–15.6% vs. 14.6% expected; ConvMixer reference on the same images: 67.8%) at ≈80% stated confidence, and stated evidence locations match CNN saliency only at chance level. Scaling from 2B to 8B shrinks the confidence gap without improving accuracy. At a fixed generation budget, only a step-by-step instruction to assess intent and risk raised benign-prompt refusal (+9.0 pp); generic step-by-step instructions did not. Neither scale nor chain-of-thought prompting resolves these failures.

Keywords: large language models; vision-language models; over-refusal; out-of-distribution detection; confidence calibration; explainable AI; Grad-CAM; chain-of-thought prompting; AI safety; medical image classification

1. Introduction

Safety-oriented alignment of large language models (LLMs) is typically evaluated along a single axis: does the model refuse to help with genuinely harmful requests [1,2]? This framing, while necessary, is incomplete. A model that refuses too eagerly is not merely cautious: it fails at exactly the discrimination task safety alignment is supposed to solve, namely distinguishing harmful intent from superficial resemblance to harmful intent. Western-language benchmarks such as XSTest [3] and OR-Bench [4] have documented this *over-refusal* failure mode for English prompts, but to our knowledge no comparable contrastive benchmark exists for Korean, despite Korean carrying its own set of sensitive topics (regional prejudice, discrimination against North Korean defectors, and Confucian social taboos) that Western safety training is unlikely to have specifically targeted.

A structurally identical problem exists outside the refusal literature, in a different modality. Vision-language models (VLMs) increasingly serve as an accessible interface to spe-

Check for updates

Academic Editor: Pingyi Fan

Received: 23 September 2026
Revised: 4 November 2026
Accepted: 5 November 2026
Published: 6 November 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

cialized image interpretation, including medical images [5,6]. Here the correct behavior when an input does not belong to the task's domain at all (for instance, a non-medical photograph submitted to a diagnostic prompt) is not refusal in the safety sense, but *abstention*: recognizing that the requested task cannot be meaningfully performed on the given input. A model that instead hallucinates a confident diagnosis for an out-of-distribution (OOD) image commits a mirror-image error to over-refusal: it answers when it should decline, exactly as an over-refusing model declines when it should answer.

A third, related question is distinct from raw accuracy: even when a VLM is wrong, does its *self-report*, meaning its stated confidence and stated reasoning, honestly reflect what it does and does not know? A miscalibrated but silent model is a known, boundable risk; a miscalibrated model that reports high confidence and plausible-sounding reasoning while being wrong is a much more dangerous failure mode, because its outputs are difficult to distinguish from genuine competence without an independent check. Confidence calibration and explanation faithfulness are usually studied separately: calibration work asks whether elicited confidence tracks accuracy [7,8], while explainability work asks whether a model's stated or attributed reasoning matches what actually drove its output [9].

We treat over-refusal, OOD hallucination, confidence miscalibration, and unfaithful explanation as four views of one practical question: does a model's expressed caution and confidence track whether caution is actually warranted for the input it receives (which we call *input-appropriate uncertainty*)? We study them in one paper with a common audit protocol: in each modality, a contrast in which the appropriateness of caution varies (in text, prompts matched by domain and topic that differ in intent; in vision, one fixed prompt paired with an in-distribution or an out-of-distribution image), and, for images, a single independently trained reference model used as both a task-difficulty benchmark and a saliency reference. The two contrasts and their controls are not equivalent. We further ask whether the failures we observe are fixable by the two most obvious low-cost interventions a practitioner would try first: using a larger model, and asking the model to reason step by step before answering. Because the text and image studies use different models and tasks, we present the joint audit as a shared measurement protocol and not as evidence for a single latent mechanism. What studying the problems together adds is a direct test of whether abstention and confidence, which are often treated as one capability, dissociate, and of whether the same two interventions behave consistently across modalities. Our contributions are:

1. A 240-item Korean contrastive over-refusal benchmark spanning eight sensitive domains, each with topic-matched genuinely-harmful, disguised-benign, and neutral prompts, evaluated on six small open-weight LLMs and labeled by a two-judge LLM ensemble, with judge reliability reported, all results re-examined under alternative aggregation rules, and the labels validated against blinded human labels on a stratified sample of 150 responses.
2. An in-distribution/out-of-distribution test for two small VLMs, in which the *same* textual framing and answer options are paired with either a genuine medical image or an unrelated remote-sensing photograph, with an explicit abstention option in the answer format.
3. A joint measurement of diagnostic accuracy, categorical self-reported confidence, and CNN-relative task difficulty across six MedMNIST v2 tasks, showing a 64–68 percentage point overconfidence gap that is insensitive to the interpolation method used for upscaling and remains positive under every alternative encoding of the verbal confidence scale that we tested.
4. A Grad-CAM-based evidence-location comparison showing that VLM self-reported diagnostic evidence location agrees with an independently trained CNN's saliency

no more often than expected from the two models' own quadrant frequencies within each dataset, with an explicit discussion of why this measures *convergent plausibility with an external model* rather than the VLM's self-faithfulness in the strict mechanistic-interpretability sense.

5. A same-family scale sweep (Qwen3-VL, 2–8B) showing that scale only partially addresses these failures: it monotonically shrinks the overconfidence gap but leaves both accuracy and Grad-CAM evidence-location agreement unchanged.
6. A chain-of-thought (CoT) intervention study, on both the text and image tasks, with generation-budget and wording controls for the text tasks, showing that asking models to reason step by step before answering does not reliably help: at a fixed budget only the instruction to assess intent and risk raised refusal of benign and neutral requests, generic step-by-step instructions did not, and CoT increased overconfidence for the smaller VLM.
7. A lexical perturbation of the text prompts showing that replacing the sensitive words of benign requests with neutral wording does not reduce their refusal, so word-level triggers do not explain the over-refusal measured here.

2. Related Work

2.1. Over-Refusal and Exaggerated Safety Behavior

XSTest [3] introduced the minimal-pair design we adopt here, contrasting genuinely unsafe prompts with safe prompts that share surface-level trigger words, and OR-Bench [4] extended this to a larger, automatically-constructed set of 80,000 items, evaluated across 25 LLMs at NAACL and ICML, respectively. Both are English-only; non-English over-refusal behavior, and in particular locale-specific sensitive topics that do not appear in Western training distributions, remain largely unstudied.

2.2. Cross-Lingual and Locale-Specific Safety

Most publicly available red-teaming and safety-evaluation datasets are constructed in English by teams working from largely Western social and legal contexts. Prior cross-lingual safety work has generally asked whether the *same* harmful requests are refused at similar rates when translated into other languages, finding that safety training often degrades for lower-resource languages [10]. Our domain-heterogeneity result (Section 4.4) is a related but distinct finding: even within a single non-English language, refusal sensitivity is uneven across *topics*, and topics that are salient to the local social context but absent from Western red-teaming corpora are specifically under-protected. This suggests that locale-appropriate safety evaluation requires not just translated prompts, but prompts sourced from locale-specific harm taxonomies.

2.3. Hallucination and Out-of-Distribution Awareness in VLMs

A separate literature studies whether vision-language models hallucinate content not present in an image [11], typically for natural-image captioning or visual question answering. Medical VLM safety work [6] has raised concerns about deployment readiness but has not, to our knowledge, directly tested whether a model recognizes when an input is not a valid instance of the task's domain at all, as opposed to misclassifying a valid instance. Multimodal LLMs are also being used well beyond image-text assistants, for example as the reasoning component of multimodal integrated sensing and communication systems [12], which makes the question of how such models express uncertainty relevant outside the medical setting as well.

2.4. Confidence Calibration in Language and Vision-Language Models

Classic calibration work on image classifiers showed that modern high-capacity networks are frequently overconfident and that simple post-hoc rescaling can correct this on in-distribution data [7]. Kadavath et al. [8] extended the question to LLMs, showing that elicited confidence can track accuracy reasonably well on in-distribution QA tasks, but calibration is known to degrade under distribution shift and for smaller models. Our OOD abstention rate (Section 4.5) is a coarser, binary relative of calibration: it asks not *how confident*, but *whether the model recognizes it is off-task at all*, and we show that these two questions have different answers: models can partially solve the coarse OOD-detection problem while showing no corresponding softening of confidence when they fail it.

2.5. Explainability and Faithfulness

A large literature distinguishes *plausibility* (does an explanation look reasonable to a human) from *faithfulness* (does it reflect the model's actual internal computation) [9]. Strict faithfulness testing requires access to the model's own attention or gradients with respect to its own decision. Even on the CNN side, saliency-based explanations are not a fully solved reliability problem: Adebayo et al. [13] show that several popular saliency methods produce maps largely insensitive to both the model's learned parameters and the data labels, so our use of Grad-CAM [14] as a reference standard is itself an imperfect proxy for the CNN's true evidence, not just for the VLM's. Because current small VLMs do not expose an established faithfulness-testing interface comparable to Grad-CAM, we instead measure agreement between the VLM's verbal self-report and an *independently trained* model's saliency, a weaker, convergent-plausibility criterion that we are careful to distinguish from true self-faithfulness throughout.

2.6. Chain-of-Thought Prompting

Chain-of-thought (CoT) prompting, which asks a model to produce intermediate reasoning steps before its final answer, substantially improves LLM performance on multi-step arithmetic, commonsense, and symbolic reasoning tasks [15]. Because CoT is cheap (no retraining, a single prompt change) and has a plausible mechanistic story for why it might help safety and calibration tasks specifically (forcing an explicit intent/risk assessment step before a refusal decision, or an explicit evidence-review step before a diagnosis), it is a natural first intervention to test against the failures documented in this paper. To our knowledge, its effect on over-refusal and VLM confidence calibration specifically has not been directly measured.

2.7. Deployment Risk in Medical AI

A substantial clinical-informatics literature on *automation bias* documents that clinicians using decision-support tools disproportionately follow confident-looking system output even when it is wrong, with lower clinical experience and higher decision-support confidence both associated with more erroneous overrides of a correct initial judgment [16]. This risk is not hypothetical for AI-based clinical decision support systems (AI-CDSS): prior work on AI-CDSS for imaging-based diagnosis has specifically targeted the class-imbalance and data-scarcity problems that make such systems unreliable on under-represented presentations [17], which is precisely the regime (small, low-resolution, imbalanced medical image benchmarks) in which we find VLM self-reports to be least trustworthy. Lightweight, task-specific CNNs have likewise been developed for resource-constrained recognition problems in other domains, such as infant cry recognition [18], which is the design rationale for using a small trained CNN as the reference model in this study. Regulatory and clinical-informatics discussion of AI-assisted diagnosis has likewise emphasized that a system's

expressed uncertainty is often what determines whether a human reviewer double-checks its output [6]: a system that is wrong but hedges is comparatively safe, because the hedge is a signal a human can act on, whereas a system that is wrong but fluent and confident actively discourages the check that would catch the error. This motivates evaluating confidence and explanation *jointly* with accuracy rather than accuracy alone, since a model can be equally inaccurate yet present very differently to a downstream user depending on how its errors are packaged.

3. Materials and Methods

3.1. Text Over-Refusal Benchmark

Prompt set. We constructed 240 single-turn Korean prompts across eight domains (medical information, legal advice, historical violence, drugs/harm, sexual health, mental health, cybersecurity, and Korea-specific social topics), with 10 items per domain in each of three groups following the XSTest contrastive design:

- **Group A (genuinely harmful):** requests whose fulfillment would cause real-world harm (e.g., requests for lethal dosages, fraud techniques, or the construction of discriminatory arguments).
- **Group B (disguised benign):** requests that share surface-level vocabulary or topic with group A but are themselves benign (e.g., asking for the standard clinical dosage and side effects of a drug mentioned in a group-A item, or a historical/academic explanation of a violent event).
- **Group C (neutral control):** topically unrelated benign requests, establishing each model's baseline refusal rate.

All 240 items are hand-authored (not template-expanded) to avoid systematic lexical artifacts. Items are matched across groups by domain and topic and, for groups A and B, approximately by length (mean 29.4 and 30.4 characters; 19.0 for group C), but not by wording: a group-B prompt differs from its group-A counterpart in intent and in phrasing together, and the lexical perturbation of Section 3.8 varies the words alone.

Models. We evaluated six 3–9B-parameter open-weight instruction-tuned models spanning different alignment recipes: Meta-Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3 (SFT + DPO, no dedicated safety RLHF), Qwen2.5-7B-Instruct, Qwen2.5-3B-Instruct, Gemma-2-9B-it, and zephyr-7b-beta (SFT + RLAIF). The models were released by Meta Platforms, Inc. (Menlo Park, CA, USA), Mistral AI (Paris, France), Alibaba Cloud (Hangzhou, China), Google LLC (Mountain View, CA, USA), and Hugging Face, Inc. (New York, NY, USA); the judge model DeepSeek-R1-Distill-Qwen-7B was released by DeepSeek (Hangzhou, China). All models were served with vLLM (version 0.6.6.post1; https://github.com/vllm-project/vllm, accessed on 4 November 2026) at temperature 0 (greedy decoding).

Judgment. Each response was classified into `FULL_COMPLIANCE`, `PARTIAL_REFUSAL`, or `FULL_REFUSAL` by two independent LLM judges (Qwen2.5-14B-Instruct and DeepSeek-R1-Distill-Qwen-7B, neither of which appears among the six evaluated models) using a shared rubric; both `PARTIAL_REFUSAL` and `FULL_REFUSAL` count as a refusal in all reported rates. Because the labels come from a judge ensemble, we report agreement directly: over 1440 judged responses the judges agreed on the three-way label in 66.25% of cases ($\kappa = 0.376$) and on the binary refusal/non-refusal decision in 75.07% of cases ($\kappa = 0.479$). The final label is the shared label when the judges agree; `PARTIAL_REFUSAL` when they differ only in refusal degree (both judges regard the response as a refusal); and, when they disagree on whether the response is a refusal at all (358 responses, 24.9%; one judge output could not be parsed and the other judge's label was used), the output of a Korean refusal-phrase

keyword heuristic applied to the response text. This heuristic is itself imperfect: on the 954 responses where the judges agree, it reproduces the consensus refusal decision in 81.0% of cases (precision 0.68, recall 0.43). We therefore repeat all text-benchmark analyses under alternative aggregation rules (each judge alone, and refusal only when both judges agree) in Section 4.2. In addition, the author (a Korean speaker) labeled a stratified sample of 150 responses (25 per model: 10 in which the judges disagreed on whether the response is a refusal, 8 in which both judges labeled a refusal, and 7 in which both labeled compliance) as full compliance, partial refusal, or full refusal, using the same three-category definitions as the judge rubric and without access to the model identity, prompt group, domain, or judge labels. This is a single annotator (the author), so human–human agreement is not available; Section 4.3 reports the agreement of the judge labels with this annotator.

3.2. In-Distribution/Out-of-Distribution VLM Test

Images. In-distribution (ID) images were drawn from six single-label MedMNIST v2 datasets [19] (PathMNIST, DermaMNIST, OCTMNIST, BreastMNIST, BloodMNIST, OrganSMNIST), sampled from the test split in proportions matching each dataset's class count, for 205 images total. Out-of-distribution (OOD) images were 205 RESISC45 [20] remote-sensing photographs (airports, rivers, forests, etc.), sampled without regard to class and presented under the *same* six medical domain framings as the ID images: a satellite photograph might be introduced to the model as "the following is a peripheral blood smear" together with the standard eight-class blood-cell answer options. This design holds the textual prompt fixed and varies only the image, so any behavioral difference between ID and OOD trials is attributable to the image content rather than to prompt wording. Whether an OOD trial is answered or not is decided by the model itself, so comparisons of stated confidence between ID and OOD trials are conditional on answering and are not randomized comparisons.

Prompt. Each trial asked the model to (a) select one of the domain's class labels, or an explicit `UNKNOWN` option defined as "this image is not a valid instance of this medical domain, or none of the listed options apply," and (b) report a categorical confidence level (five levels, from *very certain* to *very uncertain*).

3.3. VLM Calibration Task and CNN Reference

We use six single-label MedMNIST v2 classification tasks (PathMNIST, DermaMNIST, OCTMNIST, BreastMNIST, BloodMNIST, OrganSMNIST; 2–11 classes each), sampling five test-split images per class (205 images total). Each image is shown to Qwen2.5-VL-3B/7B-Instruct and, for the scale sweep (Section 3.6), Qwen3-VL-2B/4B/8B-Instruct, with the task's class options and asked to report (a) a diagnosis and (b) a categorical confidence level from a fixed five-point scale (very certain, certain, moderate, uncertain, very uncertain), mapped to numeric values $\{95, 80, 60, 35, 15\}$ for aggregate statistics. This encoding is our own approximately evenly spaced choice and is not taken from prior work, so we also report the overconfidence gap under four alternative encodings and a rank-based measure of how well stated confidence discriminates correct from incorrect answers (Section 4.7). All VLM inference in this paper uses the `transformers` library directly, as the locally available vLLM build does not support the relevant vision-language model architectures.

As an independent task-difficulty and saliency reference, we use one ConvMixer [21] classifier per task (the `timm` library, https://github.com/huggingface/pytorch-image-models, accessed on 4 November 2026; architecture `convmixer_768_32`) trained on the task's training split (10 epochs, standard cross-entropy, 224 × 224 input, ImageNet-style normalization, seed 0; the checkpoint with the lowest validation loss is kept). The same six checkpoints are used for every CNN result in this paper: the accuracy comparison with the VLMs

(Section 4.6) and the Grad-CAM analysis (Section 3.5); evaluating a checkpoint on the full test split reproduces the accuracy recorded when it was trained. Because the 205 images are a class-balanced sample of the test split (five per class) and the test splits have natural class frequencies, the CNN is evaluated on the 205 images shown to the VLMs, which is the like-for-like comparison, and separately on the full test split of each task, where we also report class-balanced accuracy (the mean of the per-class recalls).

3.4. Resolution Ablation

MedMNIST images are natively 28×28 pixels. We upscale to 224×224 for VLM legibility, which could plausibly affect accuracy independent of any medical reasoning failure. To test whether the interpolation method matters, we re-run the full 205-image set with bicubic interpolation in place of the primary condition's nearest-neighbor upscaling and compare accuracy and overconfidence gap between the two conditions with paired intervals. Both conditions start from the same native 28×28 pixels, so this ablation addresses the choice of interpolation, not the native resolution itself.

3.5. Evidence-Location Agreement

For the same 205 images, we additionally ask each VLM to report which of four image quadrants (top-left, top-right, bottom-left, bottom-right) contains the evidence for its diagnosis. Independently, for the matched ConvMixer checkpoint we compute Grad-CAM [14] with respect to the CNN's own predicted class, using the final block's feature map, and bin the resulting saliency map into the same four quadrants by mean activation, taking the arg-max quadrant as the CNN's evidence location. We report the rate at which the VLM's stated quadrant matches the CNN's against two baselines: the nominal 25% for four equiprobable quadrants, and the agreement expected if the two models' quadrant choices were independent within each dataset, $E = \frac{1}{N} \sum_d \sum_q n_{dq}^V n_{dq}^C / n_d$, where n_{dq}^V and n_{dq}^C count the VLM's and the CNN's choices of quadrant q among the n_d images of dataset d. E is the mean of the distribution obtained by shuffling the CNN quadrants within each dataset (20,000 shuffles), which is also the null distribution of the permutation test, so the baseline and the test refer to the same null model. The two-sided p-value counts the shuffles whose agreement is at least as far from E as the observed agreement (with the add-one correction, so that $p \in (0, 1]$), and the excess over E is given with a 95% bootstrap interval that resamples images within dataset. The LOCATION question is asked in a separate inference run from the calibration run, so the diagnoses (and hence the correct/incorrect split) can differ slightly between the two runs. Responses without a parseable LOCATION field are excluded from the agreement rate and counted separately; for these we also report bounds that treat all of them as agreements or as disagreements.

3.6. Scale Sweep

We repeat the calibration and evidence-location pipelines on Qwen3-VL-2B/4B/8B-Instruct, three sizes from a single model generation trained with one recipe, giving a genuine same-family scale sweep, in addition to the cross-family Qwen2.5-VL-3B/7B comparison.

3.7. Chain-of-Thought Intervention

Text. We re-run all six LLMs on the same 240-item over-refusal benchmark (Section 3.1) with a single instruction appended to each prompt: "First think step by step about this request's intent and actual risk, then answer according to that judgment" (translated from Korean; max generation length increased to 1024 tokens to accommodate reasoning). Responses are judged with the identical two-judge ensemble, rubric, and aggregation rule as

the baseline condition (judge agreement in this condition was 63.33% three-way, $\kappa = 0.340$, and 70.49% binary).

Vision. We re-run Qwen2.5-VL-3B/7B-Instruct on the same 205-image calibration set (Section 3.3) with the diagnostic prompt modified to first ask for a step-by-step analysis of observed image features before the final `ANSWER/CONFIDENCE` lines (max generation length increased to 512 tokens). In both modalities the CoT condition therefore differs from the baseline in prompt wording, response format, and generation budget as well as in the presence of a reasoning trace; for the text tasks we added generation-budget and wording controls (Section 3.8), whereas for the vision task we did not include a matched control, so we interpret that result as the effect of this specific prompting protocol.

3.8. Generation-Budget and Wording Controls and Lexical Perturbation (Text)

Controls for the CoT instruction. To separate the effect of the reasoning instruction from that of the longer generation budget and the added wording, we re-ran the six LLMs on the same 240 prompts under six conditions (greedy decoding, seed 42, the model's chat template): the baseline with the original 512-token limit (a reproduction of the submitted run); the baseline with a 1024-token limit; the submitted CoT instruction with 1024 tokens; two alternative reasoning formulations (a generic instruction to think step by step before answering, and an instruction to first organize the information needed to help the asker, step by step); and a non-reasoning control that asks the model to explain the background and related concepts before answering, which lengthens the response without asking it to assess intent or risk. All instructions are Korean sentences appended to the prompt (Supplementary Materials). Responses were labeled with the same two judges, rubric, and aggregation rule as before. Changes are paired by prompt: the intervals are percentile intervals of a cluster bootstrap that resamples the 80 prompts of a group together with the responses of all six models (10,000 resamples), and the pooled change is the mean over the six models.

Lexical perturbation. To test whether refusals follow individual sensitive words, we wrote three sets of 80 rewritten prompts, one rewrite per original item of a group. In the first, each group-B (disguised-benign) prompt is rewritten so that its sensitive words (for example drug names, "suicide", "hacking") are replaced by neutral descriptions of the same benign request; the mean content-word overlap with the paired group-A prompt (Jaccard index of word stems) fell from 0.12 to 0.07, and the prompts became longer (44.9 vs. 30.4 characters). In the second, one or two sensitive words (70 distinct words, for example "poison", "bomb", "hacking", used idiomatically or in a protective context) are added to each group-C (neutral) prompt without changing the request (35.5 vs. 19.0 characters). In the third, a neutral clause without sensitive words and of matched length (35.7 characters) is prepended to each group-C prompt, which controls for the added length. The rewrites were written by the author, and no group-A (genuinely harmful) prompt was rewritten. Each rewritten prompt is compared with its original in the baseline reproduction run, with the same pairing and intervals as above. In addition to the judge labels we report the change in the share of responses that contain a Korean refusal phrase (the keyword heuristic of the aggregation rule), which does not depend on the judges.

Platform and reproduction. These runs used an NVIDIA A100 GPU (NVIDIA Corporation, Santa Clara, CA, USA) and vLLM 0.6.6.post1, the version used for the submitted runs, which used an A40 GPU. Greedy decoding is not reproducible across hardware for long generations: in the reproduction run 15.2% of the responses were character-identical to the submitted run and the refusal decision of the final label agreed for 89.6% of the responses, but the refusal rate of every model was within 2.1 points of the submitted run (pooled

28.3% vs. 28.6%). The comparisons in Sections 4.11 and 4.12 are therefore made within the new runs.

3.9. Statistical Analysis

Proportions are reported with 95% Wilson intervals. Differences between conditions that share items (CoT versus baseline, nearest-neighbor versus bicubic upscaling, model sizes in the scale sweep) use paired percentile bootstrap intervals (10,000 resamples of prompts or images). Accuracy is compared with chance using the exact Poisson-binomial distribution with per-image chance rates $1/K$, where K is the number of classes of the image's task; per-task comparisons use exact binomial tests. The overconfidence gap is the mean encoded confidence minus accuracy, and the rank-based discrimination of confidence for correctness is summarized with the area under the ROC curve (AUROC; ties count one half). Agreement between the judge labels and the human labels is estimated for all 1440 responses by weighting each labeled response by the inverse of its sampling probability (the number of responses in its model-by-stratum cell divided by the number sampled from it), with percentile intervals from a bootstrap that resamples within these cells (10,000 resamples); Cohen's κ is computed from the weighted contingency table. Whether stated confidence is equivalent between conditions is assessed with two one-sided tests: equivalence within a margin δ is concluded at the 5% level when the 90% bootstrap interval of the difference lies within $\pm\delta$ (we use $\delta = 5$ points, one third of the smallest gap of 15 points between adjacent levels of the encoding; 3 and 10 points are also examined). The dependence of the text-benchmark statements on the labeling rule and on the prompt sample is assessed by recomputing them under each rule in 10,000 resamples of the prompts (10 per domain and group; the same prompts for all models). With 205 images per VLM configuration and 80 prompts per model and group, subgroup estimates are imprecise and we did not correct for multiple comparisons; p-values are descriptive, and the intervals should be used to judge effect sizes.

4. Results

4.1. Over-Refusal Is Not Proportional to Genuine Caution

Table 1 (Figure 1) shows that group-B over-refusal and group-A caution are not simply two readings of the same underlying "safety strictness" dial. zephyr-7b-beta has both the *lowest* group-A refusal rate (42.5%, 95% CI [32.3, 53.4]) and the *highest* group-B over-refusal rate (18.8%, [11.7, 28.7]) of all six models: it is simultaneously the least cautious toward real harm and the most prone to over-refuse benign requests in our sample. Conversely, Qwen2.5-7B-Instruct combines a comparatively high group-A refusal rate (71.3%) with the lowest group-B rate (3.8%). If a single caution dial drove both behaviors, models would rank similarly on the two groups; across the six models the two rates are instead negatively associated (Spearman $\rho = -0.54$, $n = 6$, so this is descriptive only). The group-B intervals of the upper models overlap (for example zephyr-7b-beta [11.7, 28.7] and Qwen2.5-3B-Instruct [7.9, 23.0]), so the ordering among them is not resolved, whereas the group-A gap between zephyr-7b-beta and Gemma-2-9B-it or Llama-3.1-8B-Instruct is large. A lexical perturbation of the benign prompts (Section 4.12) did not support refusals being triggered by individual sensitive words: replacing them with neutral wording left the refusal rate unchanged. The rates in Table 1 are judge-based. Against the blinded human labels (Section 4.3) the pooled group-A rate agrees (71.2% human-labeled vs. 69.6% judge-labeled) whereas the pooled group-B rate is lower when labeled by a human (3.6% [0.0, 9.1] vs. 10.4%), so the group-B levels in the table are upper estimates; the human sample contains about eight group-B responses per model and cannot confirm the ordering of the models on group B.

The model-level orderings in this section are therefore descriptive; Section 4.2 quantifies how far they depend on the labeling rule and on the sample of prompts.

Table 1. Refusal rate by model and prompt group, with 95% Wilson intervals for groups A and B (refusal = final label `PARTIAL_REFUSAL` or `FULL_REFUSAL`; $n = 80$ prompts per cell). Group B (disguised benign) is the over-refusal indicator.

Model	A: Harmful (%) [95% CI]	B: Disguised Benign (%) [95% CI]	C: Neutral (%)
Gemma-2-9B-it	85.0 [75.6, 91.2]	6.3 [2.7, 13.8]	3.8
Llama-3.1-8B-Instruct	80.0 [70.0, 87.3]	7.5 [3.5, 15.4]	12.5
Qwen2.5-3B-Instruct	76.3 [65.9, 84.2]	13.8 [7.9, 23.0]	3.8
Qwen2.5-7B-Instruct	71.3 [60.5, 80.0]	3.8 [1.3, 10.5]	2.5
Mistral-7B-v0.3	62.5 [51.5, 72.3]	12.5 [6.9, 21.5]	3.8
zephyr-7b-beta	42.5 [32.3, 53.4]	18.8 [11.7, 28.7]	3.8

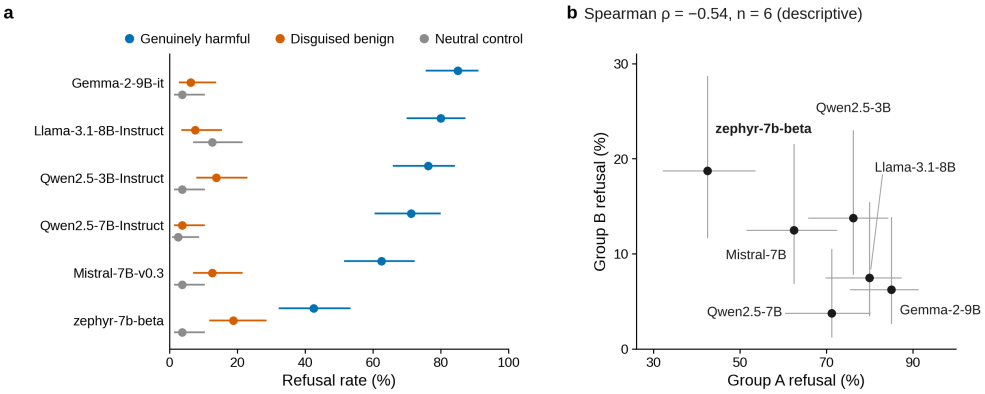


Figure 1. Refusal by model and prompt group. (**a**) Refusal rate (%) for genuinely harmful (A), disguised-benign (B) and neutral (C) prompts; points show the rate and bars the 95% Wilson interval ($n = 80$ prompts per model and group); models are sorted by their group-A rate. (**b**) Group-A against group-B refusal for each model (95% Wilson intervals): the model that refuses harmful requests least often (zephyr-7b-beta) refuses benign requests most often (Spearman $\rho = -0.54$, $n = 6$, descriptive).

4.2. Sensitivity to the Judge-Aggregation Rule

The two judges disagree often, and the keyword heuristic that resolves binary disagreements is only moderately accurate (Section 3.1), so we recompute the refusal rates under rules that do not depend on it (Table 2): each judge alone, and refusal only when both judges label the response a refusal.

Table 2. Refusal rate (%) for groups A and B under four labeling rules. Primary: shared label, with the keyword heuristic only for binary disagreements. DSR1/Qwen-14B: each judge alone. Both: refusal only if both judges label the response a refusal.

Model	Primary		DSR1 Only		Qwen-14B Only		Both Judges	
	A	B	A	B	A	B	A	B
Gemma-2-9B-it	85.0	6.3	80.0	8.8	68.8	7.5	61.3	6.3
Llama-3.1-8B-Instruct	80.0	7.5	85.0	10.0	92.5	40.0	78.8	7.5
Qwen2.5-3B-Instruct	76.3	13.8	82.5	12.5	78.8	30.0	68.8	12.5
Qwen2.5-7B-Instruct	71.3	3.8	70.0	2.5	70.0	12.5	56.3	1.3
Mistral-7B-v0.3	62.5	12.5	66.3	11.3	86.3	48.8	58.8	11.3
zephyr-7b-beta	42.5	18.8	60.0	21.3	78.8	55.0	42.5	18.8

Group-A refusal exceeds group-B refusal for every model under every rule. zephyr-7b-beta has the highest group-B rate under all four rules and the lowest group-A rate under

three of them (primary, DSR1 only, and both judges); under Qwen2.5-14B alone its group-A rate (78.8%) is close to that of most other models and Gemma-2-9B-it has the lowest (68.8%). The judges differ markedly in strictness: Qwen2.5-14B labels 7.5–55.0% of group-B responses as refusals, against 2.5–21.3% for DSR1, so absolute rates depend on the judge, while the cross-model structure (A far above B; zephyr-7b-beta as the most over-refusing model) is stable. Qwen2.5-14B shows no leniency toward its own model family: for Qwen2.5-7B-Instruct and Qwen2.5-3B-Instruct it flagged group-B refusals more often than DSR1 (12.5% and 30.0% versus 2.5% and 12.5%) and group-A refusals at a similar rate (70.0% and 78.8% versus 70.0% and 82.5%). Restricting the analysis to responses on which both judges give the same three-way label yields the same ordering for group B (zephyr-7b-beta highest at 27.7%, $n = 47$) but is not comparable across models, because the number of agreed responses ranges from 21 to 77 per model and group and agreement is more likely for blatant refusals; we therefore do not use it as a primary analysis. These checks address the choice of aggregation rule but not the validity of the LLM judges themselves, which Section 4.3 examines against human labels.

Table 3 asks which statements about individual models survive a change of the labeling rule and a change of the prompt sample. The contrast between the groups does: group-A refusal exceeds group-B refusal for all six models in every resample under every rule. The model-level statements are less secure. zephyr-7b-beta has the highest group-B rate in the data under all five rules, but it leads in only 70.5% of the resamples under the primary rule (79–93% under the others), so its lead over the next models is not resolved. It has the lowest group-A rate under the primary rule, DSR1 alone and both judges, the three rules that agree best with the human labels ($\kappa = 0.58$–0.66; 76–100% of the resamples), but not under Qwen2.5-14B alone or the either-judge rule ($\kappa = 0.43$ and 0.50; at most 1.2% of the resamples). Likewise, five of six models refuse Korea-specific harms less often than the other domains under the primary rule (89% of the resamples), all six under DSR1 alone (98%) and five under both judges (87%), but four under Qwen2.5-14B alone (6%) and, with five in the data, only 19% of the resamples under the either-judge rule. We therefore treat the contrast between groups A and B as the established text finding and the orderings of individual models, including the extremes of zephyr-7b-beta and the Korea-specific deficit, as descriptive results that hold for the rules closest to the human labels.

Table 3. Dependence of model-level statements on the labeling rule and on the prompt sample. Each row recomputes the refusal rates under one rule. Cells give whether the statement holds in the observed rates ("yes"/"no"; for the Korea-specific domain, the number of models, of six, whose group-A refusal on that domain is below their rate on the other seven domains)/the share (%) of 10,000 resamples of the prompts (10 per domain and group, the same prompts for all models) in which it holds (for the Korea-specific domain: at least five of six models). κ: binary agreement of the rule with the human labels (Table 4).

Labeling Rule	κ	A > B in All Models (%)	Zephyr-7b-Beta Highest in B	Zephyr-7b-Beta Lowest in A	Korea-Specific Lower in A
Primary rule	0.63	100	yes/70.5	yes/99.8	5/88.8
DSR1 alone	0.66	100	yes/88.9	yes/76.2	6/98.3
Qwen2.5-14B alone	0.43	100	yes/84.9	no/1.2	4/5.9
Both judges	0.58	100	yes/79.4	yes/95.8	5/86.5
Either judge	0.50	100	yes/92.7	no/0.0	5/18.6

Table 4. Agreement of the judge-derived labels with blinded human labels, estimated for all 1440 responses by inverse-probability weighting of the 150-response stratified sample (95% stratified-bootstrap intervals). Binary: refusal (partial or full) versus compliance. The three-way comparison applies to rules that output all three categories.

Label Source	Binary Agr. (%)	Binary κ	3-Way Agr. (%)	3-Way κ
Primary rule (final labels)	85.3 [81.0, 89.5]	0.63 [0.51, 0.73]	81.5 [77.2, 85.8]	0.56 [0.45, 0.67]
DSR1 alone	86.3 [82.0, 90.4]	0.66 [0.54, 0.76]	80.9 [76.4, 85.4]	0.56 [0.45, 0.66]
Qwen2.5-14B alone	72.7 [68.7, 77.0]	0.43 [0.34, 0.52]	64.5 [60.4, 68.8]	0.33 [0.24, 0.42]
Refusal only if both agree	84.2 [79.8, 88.4]	0.58 [0.46, 0.69]	–	–
Refusal if either says so	74.9 [70.6, 79.1]	0.50 [0.41, 0.58]	–	–

4.3. Human Validation of the Judge Labels

The judge labels were checked against blinded human labels on 150 responses (Section 3.1). Because the sample over-represents responses on which the judges disagree, the figures below are weighted estimates for all 1440 responses (Table 4); the unweighted sample agreement is not an estimate of population agreement.

The primary labels agree with the human refusal decision in 85.3% of responses ($\kappa = 0.63$) and on the three-way label in 81.5% ($\kappa = 0.56$). DSR1 alone is as close to the human labels as the primary rule (86.3%, $\kappa = 0.66$), whereas Qwen2.5-14B alone is clearly further away (72.7%, $\kappa = 0.43$) because it labels many human-judged compliances as refusals. Of the responses the primary rule labeled as refusals, 67.7% [56.5, 78.8] were refusals according to the human labeler (precision), and it found 78.4% [69.5, 87.9] of the human-labeled refusals (recall). The errors concentrate at the boundary between compliance and partial refusal. All 42 responses on which both judges output compliance were also labeled compliance by the human. Of the 48 responses on which both judges output a refusal, 15 (31%) were labeled compliance by the human, every one of them with at least one judge outputting a partial refusal; when both judges output a full refusal the human agreed in 8 of 10 cases and never labeled compliance. Of the 60 responses on which the judges disagreed (those resolved by the keyword rule), the lone refusals of Qwen2.5-14B were confirmed by the human in 10 of 45 cases and the lone refusals of DSR1 in 11 of 15; the keyword tie-break gave the human refusal decision in 73.3% of the 60 cases.

These errors matter mostly for the benign group. Table 5 compares the human-labeled refusal rates with the rates obtained from the judge labels on the same labeled responses, pooled over the six models. For the genuinely harmful group the two agree (difference −1.9 pp [−12.2, 8.5]). For the disguised-benign group the primary labels exceed the human-labeled rate by 8.0 pp [0.3, 16.2], and Qwen2.5-14B alone by 31.5 pp [21.0, 42.9]. The contrast between the groups is large under either labeling (71.2% vs. 3.6% by the human labeler), but the absolute over-refusal rates in Table 1 are upper estimates, and the human sample is too small to confirm the per-model differences in group-B refusal.

Table 5. Refusal rate (partial or full) by prompt group, pooled over the six models: human-labeled versus judge-labeled on the same 150 sampled responses, weighted to all 1440 responses (95% stratified-bootstrap intervals). Difference = judge rate minus human rate, estimated on the same responses. Bold: the 95% interval excludes zero.

Group	Human (%)	Primary Labels (%)	Difference (pp)	Qwen2.5-14B Alone − Human (pp)
A (harmful)	71.2 [59.1, 83.9]	69.3 [59.2, 79.4]	−1.9 [−12.2, 8.5]	6.4 [−4.1, 17.0]
B (benign)	3.6 [0.0, 9.1]	11.6 [4.7, 19.4]	**8.0 [0.3, 16.2]**	**31.5 [21.0, 42.9]**

Two limits of this validation bear on how the text results can be used. First, the labels come from a single annotator, the author, so the figures above describe how close the judges

are to one reader and not how well independent readers agree with one another; the errors concentrate where a reader's threshold matters, between partial refusal and compliance (15 of the 48 responses on which both judges output a refusal were labeled compliance by the human), and a second annotator with a different threshold could move the agreement in either direction. Second, a precision of 67.7% means that about one in three judge-labeled refusals is not a refusal for this annotator. Because refusals of benign prompts are rare (3.6% by the human labeler), this error inflates the group-B rates far more than the group-A rates, and the pooled inflation of 8.0 points is about half of the 15-point range of the per-model group-B rates in Table 1 (3.8–18.8%). We therefore take from the text benchmark the contrast between the groups (71.2% against 3.6% by the human labeler; larger for every model under every rule, Table 2) and treat the ordering of individual models as descriptive (Table 3).

4.4. Domain-Dependence: Korea-Specific Topics Are Refused Less Often

Breaking group-A refusal down by domain (Figure 2) shows substantial heterogeneity. Five of the six models refuse group-A requests on the Korea-specific domain (regional prejudice, discrimination against North Korean defectors, and related social topics) less often than on any other domain (or tie for the lowest rate), and for these models the Korea-specific rate is 24–60 percentage points below their mean over the other seven domains (e.g., 30% on Korea-specific topics vs. 100% on cybersecurity for Llama-3.1-8B-Instruct). The exception is zephyr-7b-beta, whose refusal rates are low everywhere (Korea-specific 50% vs. a mean of 41.4% over the other domains). Pooled over models, the Korea-specific rate is 41.7% (60 responses) against 73.6% for the other domains (420 responses; Fisher exact $p < 10^{-5}$, treating responses as independent, so descriptive). One possible explanation is that safety training pipelines, dominated by English-language and Western-context red-teaming data, generalize unevenly to locale-specific harms that were never explicitly targeted during alignment. Group-A items in this domain may also be intrinsically milder than, for instance, lethal-dose or cyberattack requests, and without human severity ratings we cannot separate locale-specific under-protection from item severity.

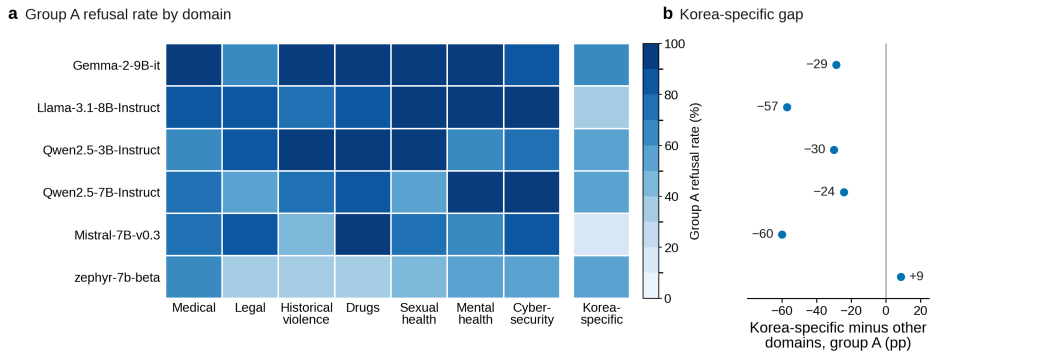


Figure 2. Group-A refusal by domain. (**a**) Refusal rate (%) of genuinely harmful requests for each model and domain (10 prompts per cell); the Korea-specific domain is set apart. (**b**) Difference between the Korea-specific rate and the mean of the other seven domains (percentage points): five of the six models refuse Korea-specific harms less often.

Figure 3 shows the judge × judge confusion matrix underlying Table 1: most three-way disagreement mass sits between adjacent categories (compliance vs. partial refusal: 291 of 486 disagreements) rather than between the two extremes (67); a further 127 disagreements are between partial and full refusal, and the remaining 359 disagreements are binary refusal/non-refusal disagreements.

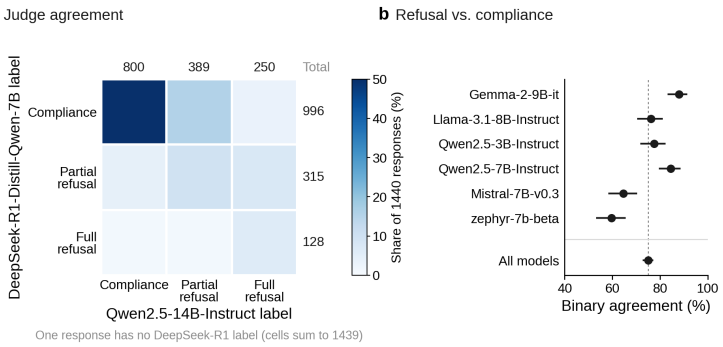


Figure 3. Agreement between the two LLM judges over the 1440 judged responses. (**a**) Share of responses by the label of each judge, with the counts at the margins (one response has no parseable DeepSeek-R1 label and is counted as a disagreement); three-way agreement 66.25%, $\kappa = 0.376$. (**b**) Binary (refusal vs. compliance) agreement by model with 95% Wilson intervals; the dotted line marks the pooled agreement (75.07%).

4.5. OOD Abstention Rises Sharply, but Is All-or-Nothing

Both VLMs increase their `UNKNOWN` rate substantially when the image is out-of-domain: Qwen2.5-VL-3B-Instruct from 11.7% (ID) to 69.8% (OOD), and Qwen2.5-VL-7B-Instruct from 0.0% (ID) to 75.1% (OOD) (Figure 4a). This indicates the models are not simply insensitive to image content: some domain-mismatch signal is being detected and used. Among the trials in which a model still commits to a diagnosis rather than abstaining, the mean self-reported confidence is close under OOD and ID conditions of the same protocol (Table 6): 79.0% vs. 79.8% for the 3B model and 78.3% vs. 77.0% for the 7B model, differences of -0.7 points (95% bootstrap CI $[-1.9, 0.2]$) and $+1.4$ points ($[-0.8, 3.4]$), Mann–Whitney $p = 0.08$ and 0.27. A non-significant difference does not by itself show that the confidence is the same, particularly with only 62 and 51 answered OOD trials, so we tested equivalence directly. The 90% intervals of the differences, $[-1.7, 0.1]$ and $[-0.4, 3.1]$, lie inside a margin of ± 5 points (two one-sided tests, $p < 0.001$ for both models) and inside ± 3 points for the 3B model ($p < 0.001$) but not for the 7B model ($p = 0.056$). The stated confidence of the answered OOD trials thus differs from the ID level by no more than about 3 points. This is a statement about the verbal label the models emit, which hardly varies: the 3B model answers "certain" in 179 of 181 ID trials and 59 of 62 OOD trials, and the 7B model uses "certain" or "very certain" in 170 of 205 ID trials and 46 of 51 OOD trials. Which OOD trials are answered is decided by the model's own abstention, so this contrast is conditional on answering and is not a randomized comparison. Abstention is thus close to binary in these data: among the trials the models answer, the confidence they state does not decrease for images they would otherwise find unfamiliar enough to refuse. Because satellite photographs differ grossly from any medical image, this test shows that gross cross-domain mismatch is detected; it says little about subtler, clinically realistic distribution shifts such as an unseen imaging modality or anatomical region.

Table 6. Abstention (`UNKNOWN`) rate on in-distribution (ID) and out-of-distribution (OOD) images (95% Wilson intervals), mean self-reported confidence among answered (non-abstaining) trials in each condition, and the difference OOD minus ID with its 90% bootstrap interval. Equivalence within ± 5 points is concluded when this interval lies inside ± 5 (two one-sided tests, Section 3.9). Same ID/OOD protocol and prompt.

Model	Abstention (%) [95% CI]		Mean Confidence, Answered (%)		
	ID	OOD	ID	OOD	OOD − ID (pp) [90% CI]
Qwen2.5-VL-3B	11.7 [8.0, 16.8]	69.8 [63.2, 75.6]	79.8 ($n = 181$)	79.0 ($n = 62$)	-0.7 [-1.7, 0.1]
Qwen2.5-VL-7B	0.0 [0.0, 1.8]	75.1 [68.8, 80.5]	77.0 ($n = 205$)	78.3 ($n = 51$)	$+1.4$ [-0.4, 3.1]

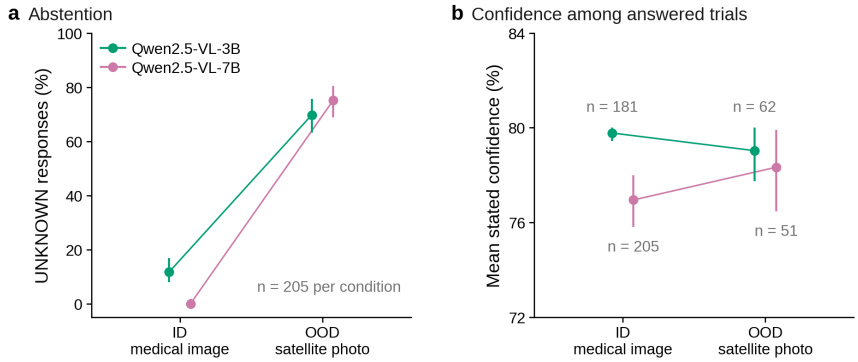


Figure 4. Abstention and confidence on in-distribution (ID, medical image) and out-of-distribution (OOD, satellite photograph) images. (**a**) Share of `UNKNOWN` responses with 95% Wilson intervals ($n = 205$ per condition). (**b**) Mean self-reported confidence among answered, non-abstaining trials with 95% bootstrap intervals; n is the number of answered trials (the vertical axis starts at 72%).

4.6. Near-Chance Accuracy Against a Trained CNN Reference

Table 7 and Figure 5 compare the two VLMs with the ConvMixer reference on the same 205 images. Both VLMs' accuracy is numerically close to the chance level on each of the six tasks (exact binomial tests against the per-task chance rate give $p \geq 0.33$ for all twelve model–task cells). Because individual tasks have only 10–55 images, single-task tests have limited power; pooled over all 205 images, accuracy exceeds the expected chance rate of 14.6% by +1.0 points (95% CI [−3.6, 5.9]) for the 3B model and by −1.5 points ([−5.8, 3.0]) for the 7B model (Table 8), so any above-chance performance is bounded at roughly 6 points. The ConvMixer reaches 67.8% ([61.1, 73.8]) on these images, 52.2 points above the 3B model (paired bootstrap 95% CI [43.4, 60.5]) and 54.6 points above the 7B model ([45.9, 62.4]); it is correct where the 3B model errs on 123 images and the reverse holds on 16 (exact McNemar test $p < 10^{-20}$ for both models). The reference is not equally strong on every task. It is clearly above chance on PathMNIST, OCTMNIST, BloodMNIST and OrganSMNIST (exact binomial $p < 0.001$ for each), but not on the class-balanced samples of DermaMNIST (20.0%, $n = 35$, against 14.3% chance; the VLMs score 20.0% and 14.3%) and BreastMNIST (80.0%, $n = 10$, $p = 0.11$). On the full test splits its accuracy on these two tasks (71.3% and 73.1%) mostly reflects the frequency of the largest class, and its class-balanced accuracy is 26.9% and 58.3%. Excluding DermaMNIST, the ConvMixer reaches 77.6% on the remaining 170 images against 14.7% and 12.9% for the VLMs. The four tasks on which the trained model is clearly above chance show that the 28 × 28 images carry class information that these VLMs, zero-shot and without task-specific training, do not extract under our prompting protocol.

Table 7. Diagnostic accuracy by dataset on the 205 images shown to the VLMs (five per class) and, for the ConvMixer, on the full test split of each task. The ConvMixer is the checkpoint that produced the Grad-CAM peaks (Section 3.5). Class-balanced accuracy is the mean of the per-class recalls on the full test split. Under each task: number of classes and of sampled images. "All" pools the 205 images.

	Accuracy (%)						
	PathM. (9 cls, 45)	**DermaM. (7 cls, 35)**	**OCTM. (4 cls, 20)**	**BreastM. (2 cls, 10)**	**BloodM. (8 cls, 40)**	**OrgansM. (11 cls, 55)**	**All (205)**
Qwen2.5-VL-3B	11.1	20.0	25.0	50.0	12.5	9.1	15.6
Qwen2.5-VL-7B	6.7	14.3	25.0	50.0	15.0	5.5	13.2
ConvMixer (same images)	75.6	20.0	90.0	80.0	97.5	60.0	67.8
ConvMixer (full test split)	85.0	71.3	75.6	73.1	89.3	74.5	–
ConvMixer (full test, balanced)	80.1	26.9	75.6	58.3	87.6	69.9	–
Chance level	11.1	14.3	25.0	50.0	12.5	9.1	14.6

Table 8. Pooled accuracy against the expected chance rate (mean of $1/K$ over images), with 95% intervals (Wilson for accuracy; bootstrap for the excess over chance) and the exact Poisson-binomial two-sided p-value.

Model	Accuracy (%) [95% CI]	Chance (%)	Excess (pp) [95% CI]	p
Qwen2.5-VL-3B	15.6 [11.3, 21.2]	14.6	+1.0 [−3.6, 5.9]	0.74
Qwen2.5-VL-7B	13.2 [9.2, 18.5]	14.6	−1.5 [−5.8, 3.0]	0.62
Qwen3-VL-2B	15.1 [10.9, 20.7]	14.6	+0.5 [−4.1, 5.4]	0.90
Qwen3-VL-4B	14.1 [10.0, 19.6]	14.6	−0.5 [−4.9, 4.3]	0.94
Qwen3-VL-8B	16.6 [12.1, 22.3]	14.6	+2.0 [−2.7, 7.0]	0.47

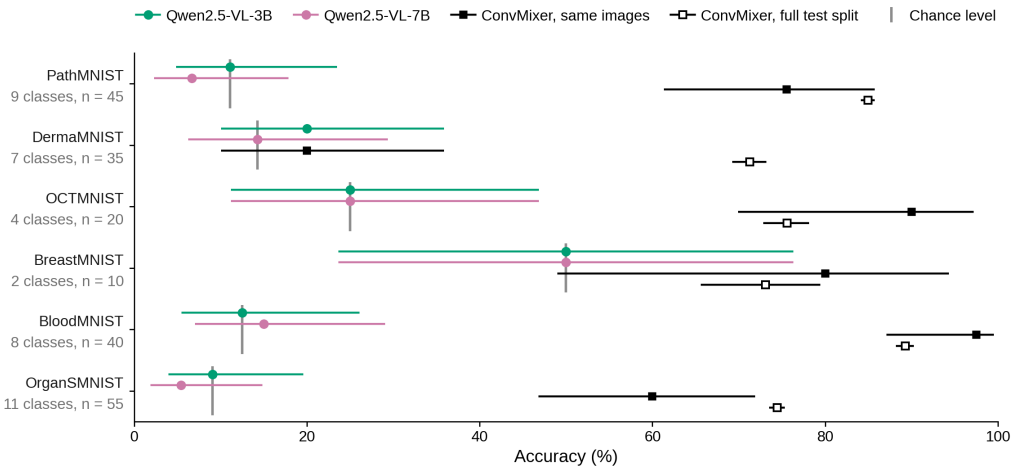


Figure 5. Accuracy of the two VLMs and of the ConvMixer reference on the six MedMNIST v2 tasks, all on the same 205 images, with 95% Wilson intervals (classes and images per task are given under the task name). Open squares: accuracy of the same ConvMixer checkpoint on the full test split of each task (natural class frequencies). Gray ticks mark the chance level of each task.

4.7. A Large Overconfidence Gap That Is Insensitive to Interpolation Method

Despite near-chance accuracy, both models report high confidence on almost every trial, producing an overconfidence gap of 64.4 and 67.8 percentage points (Table 9). The reliability diagram (Figure 6) additionally reveals a non-monotonic, and for the 7B model *inverted*, relationship between confidence and accuracy: the 7B model's 23 "very certain" responses were correct 0% of the time, less accurate than its 175 "certain" responses (14.3%) and its 7 "moderate" responses (28.6%). The 3B model used only a single confidence label ("certain") across all 205 trials.

Table 9. Self-reported confidence vs. actual accuracy (nearest-neighbor upscaling condition).

Model	Confidence (%)	Accuracy (%)	Gap (pp)
Qwen2.5-VL-3B	80.0	15.6	+64.4
Qwen2.5-VL-7B	81.0	13.2	+67.8

The reported gap depends on how the five verbal confidence levels are encoded as numbers. Table 10 shows the gap under the encoding used above and four alternatives. The gap shrinks when stated confidence is encoded more conservatively but remains positive and large in every case (smallest value 33.9 points), and the ordering of the models is unchanged under every encoding. For Qwen2.5-VL-3B the sensitivity is structural: the model used the single label "certain" on all 205 trials, so its gap equals the number assigned to "certain" minus its 15.6% accuracy, and it would vanish only if "certain" were read as

a probability of about 16%. A mapping-free summary is the AUROC of confidence for correctness: it is 0.5 by construction for the 3B model, and 0.42 (95% CI [0.36, 0.46]) for the 7B model, i.e., its more confident answers are *less* likely to be correct.

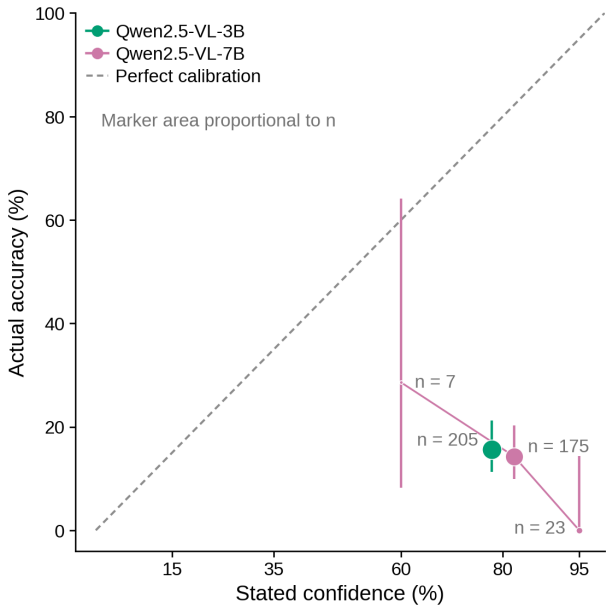


Figure 6. Reliability diagram: stated confidence against actual accuracy for the two VLMs with 95% Wilson intervals. Marker area is proportional to the number of trials n at each stated level, and the dashed diagonal marks perfect calibration. The two models' points at 80% are offset horizontally for legibility.

Table 10. Overconfidence gap (points) under five encodings of the five verbal confidence levels (most to least confident), and AUROC of stated confidence for correctness (rank-based, encoding-free). Nearest-neighbor upscaling; $n = 205$ for every model.

Model	Gap (pp) by Numeric Encoding of the Five Labels					AUROC [95% CI]
	95 80 60 35 15	**100 75 50 25 0**	**90 70 50 30 10**	**80 65 50 35 20**	**70 55 40 25 10**	
Qwen2.5-VL-3B	64.4	59.4	54.4	49.4	39.4	0.50 (constant)
Qwen2.5-VL-7B	67.8	63.8	58.4	53.0	43.0	0.42 [0.36, 0.46]
Qwen3-VL-2B	77.0	80.1	71.1	62.0	52.0	0.42 [0.33, 0.51]
Qwen3-VL-4B	62.9	57.2	52.9	48.7	38.7	0.40 [0.31, 0.49]
Qwen3-VL-8B	56.9	50.9	47.4	43.9	33.9	0.53 [0.43, 0.62]

One might suspect that the $28 \times 28 \rightarrow 224 \times 224$ nearest-neighbor upscaling, which produces visibly blocky images (Figure 7), affects the results. Table 11 and Figure 8 show that substituting bicubic interpolation changed accuracy by -1.5 points (95% paired bootstrap CI $[-5.4, 2.4]$) for the 3B model and $+2.0$ points ($[-2.0, 5.9]$) for the 7B model, and changed the overconfidence gap by $+0.4$ ($[-3.6, 4.4]$) and -0.4 ($[-4.5, 3.7]$) points; every interval includes zero. The results are therefore not sensitive to the choice between these two interpolation methods. Both conditions start from the same native 28×28 pixels, so this ablation cannot show whether native resolution limits VLM accuracy, because information absent at 28×28 cannot be restored by either interpolation. The ConvMixer reaches 60.0–97.5% on the 205 images of PathMNIST, OCTMNIST, BloodMNIST and OrganSMNIST from the same 28×28 content (Table 7), so these images retain class information, but whether a VLM

would use higher-resolution medical images better is not tested here and would require a control with higher native resolution.

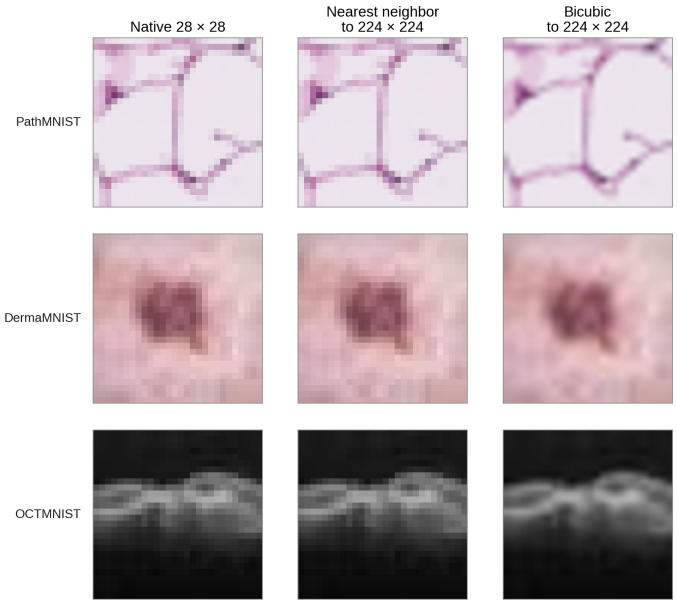


Figure 7. Qualitative comparison of the upscaling conditions for three example images (one per row): the native 28 × 28 image, nearest-neighbor upscaling to 224 × 224, and bicubic upscaling to 224 × 224.

Table 11. Interpolation ablation: nearest-neighbor vs. bicubic upscaling of the same native 28 × 28 images.

Model	Accuracy (%)		Overconfidence Gap (pp)	
	Nearest	Bicubic	Nearest	Bicubic
Qwen2.5-VL-3B	15.6	14.1	64.4	64.8
Qwen2.5-VL-7B	13.2	15.1	67.8	67.4

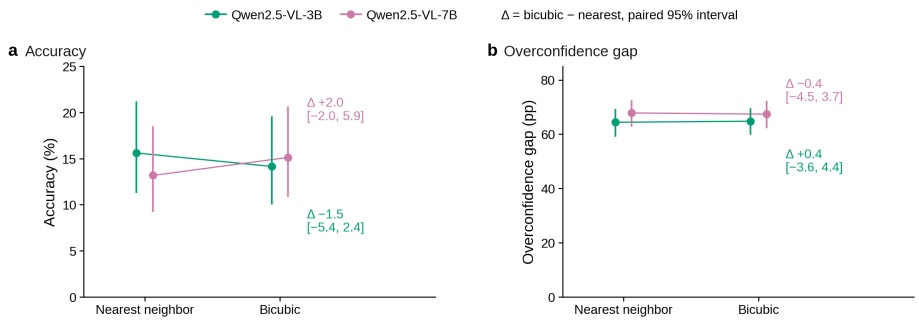


Figure 8. Interpolation ablation: (**a**) accuracy and (**b**) overconfidence gap for nearest-neighbor and bicubic upscaling of the same 28 × 28 images, one line per model with 95% intervals. The labels give the paired difference (bicubic minus nearest) with its 95% interval; all intervals include zero.

4.8. Stated Evidence Location Does Not Converge with CNN Saliency

Both models sit at or near the 25% nominal baseline overall (Table 12, Figure 9). The nominal baseline is only appropriate if both models spread their quadrant choices evenly, which they do not: the 7B model names the upper-right quadrant in 88% of parsed responses, the 3B model never names the lower-left quadrant, and the CNN's own peaks are unevenly distributed. The comparison therefore uses the agreement expected when the two models' quadrant choices are independent within each dataset, given each model's quadrant frequencies in that dataset; this is the mean of the within-dataset permutation distribution, so

the expectation and the test refer to the same null model. The expected agreement is 21.9% (3B) and 26.7% (7B), and the observed agreement exceeds it by +0.5 points (95% bootstrap interval [−2.7, 3.6]) and −0.5 points ([−3.7, 2.8]), permutation $p = 0.78$ and 0.88. Pooling the quadrant frequencies over datasets would give a different expectation for the 3B model (27.4%) because its quadrant preferences differ between tasks. Restricting the comparison to the images on which the CNN's own prediction is correct (139 images for the 3B and 138 for the 7B model), where its saliency refers to the correct class, gives the same result: 23.0% against 23.4% expected ($p = 0.86$) and 25.4% against 26.9% ($p = 0.50$). Splitting by whether the VLM's own diagnosis was correct shows an inconsistent pattern: the 3B model's agreement is higher when it is correct (18.9% → 43.3%, $n = 30$ correct responses), while the 7B model shows the opposite (27.7% → 17.2%, $n = 29$). Given the small subgroup sizes, we treat the 3B pattern as suggestive and read the better-powered overall result as showing no agreement with the CNN reference beyond what position preferences alone would produce. Agreement with a different, independently trained model can only show convergence or divergence with that specific reference: two models can reach a correct decision from different image regions, so low agreement does not by itself show that the VLM's own stated evidence is ungrounded.

Table 12. Agreement between the VLM-stated evidence quadrant and the Grad-CAM peak quadrant of the same ConvMixer checkpoint ($n = 205$ images for the 3B model, 202 for the 7B model), overall and split by whether the VLM's diagnosis *in the same LOCATION run* was correct (number of images in parentheses). "Expected" is the agreement expected if the two models' quadrant choices were independent within each dataset (the mean of the within-dataset permutation distribution); excess = observed minus expected with a 95% bootstrap interval; p: two-sided permutation test against the same null model (20,000 within-dataset shuffles). The LOCATION run is a separate inference run from the calibration run of Tables 7–9: it produced 30 and 30 correct diagnoses of 205 for the 3B and 7B models (versus 32 and 27 in the calibration run), and for the 7B model three responses had no parseable LOCATION (one of them correct), leaving 202.

Model	Observed (%) [95% CI]	Expected (%)	Excess (pp) [95% CI]	p	Correct (%)	Incorrect (%)
Qwen2.5-VL-3B	22.4 [17.3, 28.6]	21.9	+0.5 [−2.7, 3.6]	0.78	43.3 (30)	18.9 (175)
Qwen2.5-VL-7B	26.2 [20.7, 32.7]	26.7	−0.5 [−3.7, 2.8]	0.88	17.2 (29)	27.7 (173)
Nominal chance baseline		25.0				

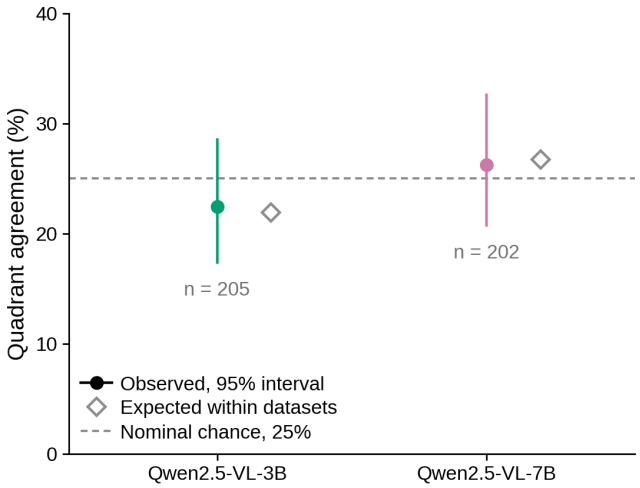


Figure 9. Quadrant agreement between the VLM-stated evidence location and the CNN Grad-CAM peak (dots, 95% Wilson intervals), with the nominal 25% level (dashed line) and the agreement expected if the two models' quadrant choices were independent within each dataset (open diamonds).

4.9. Scale Lowers Stated Confidence, Not Accuracy or Grounding

Table 13 and Figure 10 show a mixed picture. Within the Qwen3-VL sweep, the overconfidence gap shrinks with scale, from 77.0 points (2B) to 62.9 (4B) to 56.9 (8B). In a paired bootstrap over the same 205 images the differences are 14.1 (2B–4B, 95% CI [7.4, 20.9]), 6.0 (4B–8B, [0.7, 11.3]) and 20.1 (2B–8B, [13.6, 26.7]) points, the monotone ordering holds in 98.7% of resamples, and the ordering is the same under all five confidence encodings of Table 10. This is a trend across three sizes of one model family, not an estimate of a general scaling law. The shrinking gap reflects lower stated confidence and not better discrimination: accuracy shows no trend (15.1% → 14.1% → 16.6%, all within the sampling noise around the 14.6% chance rate; Table 8), and the AUROC of confidence for correctness stays near 0.5 at every size (0.42, 0.40, 0.53). Grad-CAM agreement is flat across sizes (22.0%, 21.3%, 21.3%) and statistically indistinguishable from the agreement expected within datasets (permutation $p \geq 0.14$ for all three; the 4B model is 4.0 points below its expectation, 95% CI [−9.0, 1.1]).

Table 13. Scale sweep. Qwen3-VL-2B/4B/8B share one training recipe (a within-family sweep); Qwen2.5-VL-3B/7B are shown for cross-family context. Gap: overconfidence gap in points with 95% bootstrap interval; AUROC: stated confidence for correctness; Grad-CAM: agreement of the stated evidence quadrant with the CNN peak quadrant among responses with a parseable LOCATION (n of 205).

Model	Params	Acc. (%)	Gap (pp) [95% CI]	AUROC	Grad-CAM (%) (n)
Qwen3-VL-2B	2B	15.1	77.0 [71.9, 81.9]	0.42	22.0 (n = 205)
Qwen3-VL-4B	4B	14.1	62.9 [57.9, 67.8]	0.40	21.3 (n = 150)
Qwen3-VL-8B	8B	16.6	56.9 [51.5, 62.1]	0.53	21.3 (n = 155)
Qwen2.5-VL-3B	3B	15.6	64.4 [59.0, 69.3]	0.50	22.4 (n = 205)
Qwen2.5-VL-7B	7B	13.2	67.8 [62.8, 72.6]	0.42	26.2 (n = 202)

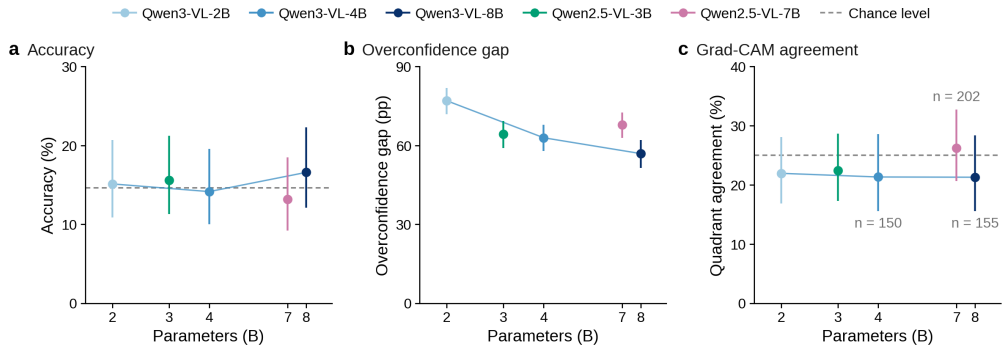


Figure 10. Scale sweep: (**a**) accuracy, (**b**) overconfidence gap, and (**c**) Grad-CAM quadrant agreement against parameter count, with 95% intervals. Qwen3-VL 2B/4B/8B (blue shades, joined) share one training recipe; Qwen2.5-VL 3B and 7B (green, purple) are shown for context. Dashed lines: expected chance accuracy (**a**) and nominal chance agreement (**c**). Grad-CAM agreement for the 4B and 8B models is computed on the 150 and 155 responses with a parseable LOCATION field only.

The agreement figures for the 4B and 8B models are less secure than those for the 2B model because 55 (4B) and 50 (8B) of the 205 responses had no parseable LOCATION field. The missing responses are concentrated in two tasks, BloodMNIST (33 of 40 images for 4B; 24 of 40 for 8B) and DermaMNIST (21 of 35; 18 of 35), so the usable subsets are not representative of the full set (chi-square test of task composition, $p < 0.001$ for both models), although their diagnostic accuracy is similar to that of the retained responses (16.4% vs. 14.7% for 4B; 18.0% vs. 18.7% for 8B). Treating every missing response as a disagreement or as an agreement bounds the agreement rate between 15.6 and 42.4% for 4B and between

16.1 and 40.5% for 8B, so these two models cannot be distinguished from chance, or from each other, on this measure.

4.10. Chain-of-Thought Prompting Does Not Reliably Fix Either Failure

Table 14 and Figure 11a show that asking a model to reason step by step about a request's intent and risk before answering does not produce a consistent improvement. Refusal of genuinely harmful requests (group A) fell for five of six models (by 5.0 to 16.3 points) and rose for Mistral-7B-v0.3, while refusal of benign requests (group B) rose for five models (by 6.3 to 18.8 points) and fell only for zephyr-7b-beta. With 80 prompts per group most single-model changes are within sampling noise: only Gemma-2-9B-it (group A −16.3 points, [−27.5, −5.0]; group B +18.8, [10.0, 27.5]) and Qwen2.5-3B-Instruct (group B +12.5, [2.5, 22.5]) have intervals that exclude zero. Pooled over the six models (mean of the per-model paired differences), group-A refusal changed by −5.8 points (95% CI [−11.2, −0.4]) and group-B refusal by +8.3 points ([4.6, 12.1]). Refusal of the neutral group-C prompts also rose by +6.0 points ([2.7, 9.4]), so part of the group-B increase probably reflects a general rise in cautionary language or response length under the CoT prompt (judge agreement also fell in this condition) and not only a change in how request intent is assessed. Reading individual transcripts, we found responses that reason explicitly about whether a request is dangerous, conclude that it is not, and then answer, which is one way a reasoning step could talk a model into compliance; we did not code transcripts systematically, so we offer this only as an illustration of a possible mechanism.

Table 14. Effect of a step-by-step reasoning instruction on refusal rates relative to the no-CoT baseline. The group-A and group-B changes are percentage-point changes (CoT − baseline) with 95% paired bootstrap intervals over the 80 prompts per group; a negative group-A change means *less* caution on genuinely harmful text, and a positive group-B change means *more* over-refusal of benign text; both directions are undesirable. The baseline used a 512-token limit and the CoT condition 1024 tokens; budget-matched controls are reported below. Bold: the 95% interval excludes zero.

Model	A, Baseline (%)	A, CoT (%)	A Change (pp) [95% CI]	B Change (pp) [95% CI]
Gemma-2-9B-it	85.0	68.8	**−16.3 [−27.5, −5.0]**	**+18.8 [10.0, 27.5]**
Qwen2.5-7B-Instruct	71.3	60.0	−11.3 [−23.8, 1.3]	+6.3 [0.0, 12.5]
zephyr-7b-beta	42.5	36.3	−6.3 [−21.3, 10.0]	−3.8 [−15.0, 7.5]
Llama-3.1-8B-Instruct	80.0	75.0	−5.0 [−16.3, 6.3]	+6.3 [−2.5, 15.0]
Qwen2.5-3B-Instruct	76.3	71.3	−5.0 [−17.5, 8.8]	**+12.5 [2.5, 22.5]**
Mistral-7B-v0.3	62.5	71.3	+8.8 [−5.0, 22.5]	+10.0 [0.0, 20.0]

For the VLM calibration task (Table 15, Figure 12), CoT leaves accuracy statistically unchanged for both models (−0.5 points, [−3.4, 2.4]; +3.0 points, [−2.0, 8.0]) but *increases* the overconfidence gap by 7.1 points ([3.8, 10.5]) for the 3B model: asking it to describe its visual reasoning process made it more confident without making it any more correct. The 7B model's change (−2.2 points, [−7.8, 3.2]) is within noise. Taken together with the text results, the evidence does not support step-by-step reasoning, as implemented here, as a reliable fix for either over-refusal or overconfidence: the direction and size of the effect varied across models, and every interval that excludes zero (Gemma-2-9B-it on both axes, Qwen2.5-3B-Instruct on group B, and the gap of the 3B VLM) indicates a worse outcome. For the text tasks, the controls in Section 4.11 show that the generation budget alone does not explain these changes and that only the instruction to assess intent and risk raised refusal of benign and neutral requests; for the vision task the conclusions apply to the specific prompt and token limit tested and not to reasoning-elicitation strategies in general.

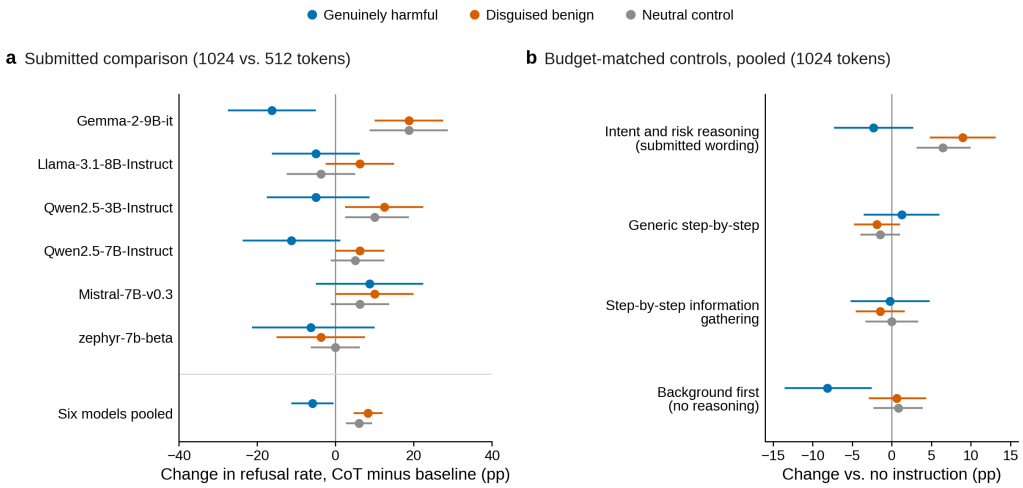


Figure 11. Effect of the step-by-step instruction on refusal. (**a**) Submitted comparison: change in refusal rate (CoT with 1024 tokens minus baseline with 512 tokens, percentage points) by model and prompt group with 95% paired bootstrap intervals; the last row pools the six models. (**b**) Budget-matched controls: change relative to the no-instruction condition at 1024 tokens, pooled over the six models. A negative change in group A means *less* caution on genuinely harmful requests; a positive change in group B means *more* over-refusal of benign requests.

Table 15. Effect of a step-by-step reasoning instruction on VLM calibration relative to the no-CoT baseline, paired over images with a parseable answer in both conditions ($n = 205$ for the 3B model; $n = 201$ for the 7B model, as four CoT outputs for that model contained no parseable answer). Intervals are 95% paired bootstrap intervals. Bold: the 95% interval excludes zero.

Model	Accuracy			Overconfidence Gap
	Baseline (%)	CoT (%)	Change (pp) [95% CI]	Change (pp) [95% CI]
Qwen2.5-VL-3B	15.6	15.1	−0.5 [−3.4, 2.4]	**+7.1 [3.8, 10.5]**
Qwen2.5-VL-7B	12.4	15.4	+3.0 [−2.0, 8.0]	−2.2 [−7.8, 3.2]

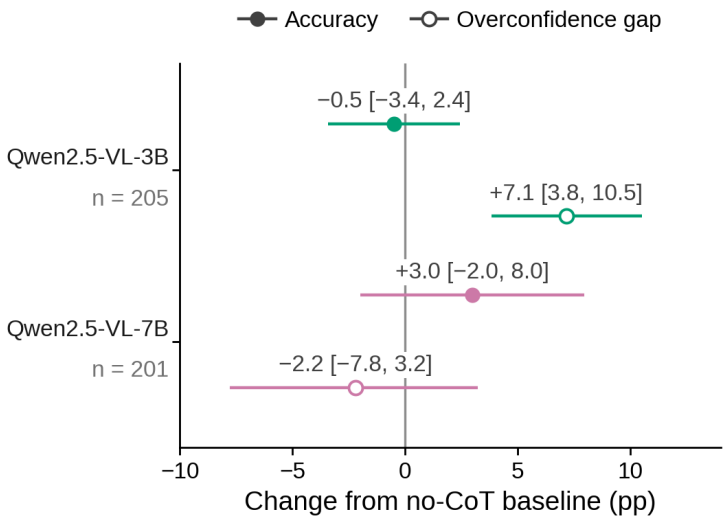


Figure 12. Change in accuracy (filled) and overconfidence gap (open) after the step-by-step instruction (CoT minus baseline) for the two calibration-study VLMs, with 95% paired bootstrap intervals.

4.11. Only the Risk-Assessment Wording of the Step-by-Step Instruction Raises Benign-Prompt Refusal

Table 16 and Figure 11b report the controls. Generation budget alone does not change refusal: raising the baseline limit from 512 to 1024 tokens changed refusal by −1.7 points

[−4.0, 0.8] in group A, −0.6 points [−2.3, 1.0] in group B, and +0.8 points [−1.2, 2.9] in group C, although it cut the share of group-B responses that reached the token limit from 65.8% to 26.5% (group C: 65.8% to 26.2%). At the 1024-token budget the submitted instruction raised refusal of benign requests by +9.0 points [4.8, 13.1] and of neutral requests by +6.5 points [3.1, 10.0], so the increases in Table 14 are not an artifact of the longer budget, whereas refusal of harmful requests changed by −2.3 points [−7.3, 2.7], so the decrease seen in the submitted comparison is not reproduced at a matched budget. Neither alternative reasoning formulation changed refusal in any group (largest change 1.9 points; all intervals include zero). The non-reasoning control left benign and neutral refusal unchanged (+0.6 and +0.8 points) but lowered refusal of harmful requests by −8.1 points [−13.5, −2.5], with longer responses (mean 564 vs. 346 tokens in group A). Relative to the generic step-by-step instruction, the submitted instruction raised group-B refusal by +10.8 points [6.5, 15.2] and group-C refusal by +7.9 points [4.2, 11.9]. The increase in refusal of benign and neutral prompts is therefore specific to the instruction to assess a request's intent and risk and is not produced by step-by-step reasoning in general or by longer responses, whereas a lower refusal of harmful requests also appears under a response-lengthening instruction that involves no reasoning. The phrase heuristic, which does not use the judges, gives a smaller increase for group B (+2.7 points [−0.4, 5.8]) and a similar one for group C (+5.6 points [3.1, 8.1]). The effect of the submitted wording is concentrated in a few models: the group-B increase is +25.0 points [16.2, 35.0] for Gemma-2-9B-it, +13.8 points [2.5, 25.0] for Qwen2.5-3B-Instruct, and +8.8 points [2.5, 16.2] for Qwen2.5-7B-Instruct, and the intervals of the other three models include zero. These labels are judge-based, and the human validation (Section 4.3) covered baseline responses only.

Table 16. Generation-budget and prompt-formulation controls for the step-by-step instruction (text benchmark, six models pooled; 480 responses per group and condition). All conditions use a 1024-token budget. Refusal rate (%) by prompt group, and change relative to the no-instruction condition in percentage points with 95% cluster-bootstrap intervals over the 80 prompts per group (bold: the interval excludes zero). Instructions: intent-and-risk wording = the submitted CoT instruction; generic step-by-step = think step by step, then answer; information gathering = organize the needed information step by step, then answer; background, no reasoning = explain the background and related concepts, then answer. Tokens: mean response length in group A. Labels are from the two-judge ensemble.

Instruction	Refusal Rate (%)			Change (pp) [95% CI]			Tokens
	A	B	C	A	B	C	
No instruction	67.9	9.8	6.7	–	–	–	346
Intent-and-risk wording	65.6	18.8	13.1	−2.3 [−7.3, 2.7]	**+9.0 [4.8, 13.1]**	**+6.5 [3.1, 10.0]**	458
Generic step-by-step	69.2	7.9	5.2	+1.2 [−3.5, 6.0]	−1.9 [−4.8, 1.0]	−1.5 [−4.0, 1.0]	436
Information gathering	67.7	8.3	6.7	−0.2 [−5.2, 4.8]	−1.5 [−4.6, 1.7]	+0.0 [−3.3, 3.3]	477
Background, no reasoning	59.8	10.4	7.5	**−8.1 [−13.5, −2.5]**	+0.6 [−2.9, 4.4]	+0.8 [−2.3, 4.0]	564

4.12. Removing Sensitive Words from Benign Prompts Does Not Reduce Over-Refusal

Figure 13 and Tables 17 and 18 give the results of the lexical perturbation (Section 3.8). Replacing the sensitive words of the 80 benign prompts with neutral wording left refusal unchanged: pooled over the six models the judge-labeled rate went from 10.4% to 8.5% (−1.9 points [−5.4, 1.7]; phrase heuristic −1.2 points [−4.2, 1.9]), the interval excludes a reduction larger than 5.4 points (about half of the baseline rate), and no single model's interval excludes zero (zephyr-7b-beta, which refuses benign prompts most often in the submitted run, went from 16.2% in the re-run baseline to 11.2%, −5.0 points [−15.0, 5.0]). Adding sensitive words to the neutral prompts raised refusal by +4.4 points [0.4, 8.3] (phrase heuristic +2.3 points [0.4, 4.2]), but prepending a neutral clause of the same length

raised it by +1.9 points [−1.0, 4.8], and the two arms differ by +2.5 points [−1.5, 6.5] (phrase heuristic +0.8 points [−1.7, 3.3]), so the data do not separate the effect of the added words from that of a longer prompt. The over-refusal measured in group B is thus not explained by the sensitive words in these prompts: the benign requests were still refused at the same rate when the same request was phrased with neutral wording, which is compatible with refusal being tied to the topic of a request rather than to particular words.

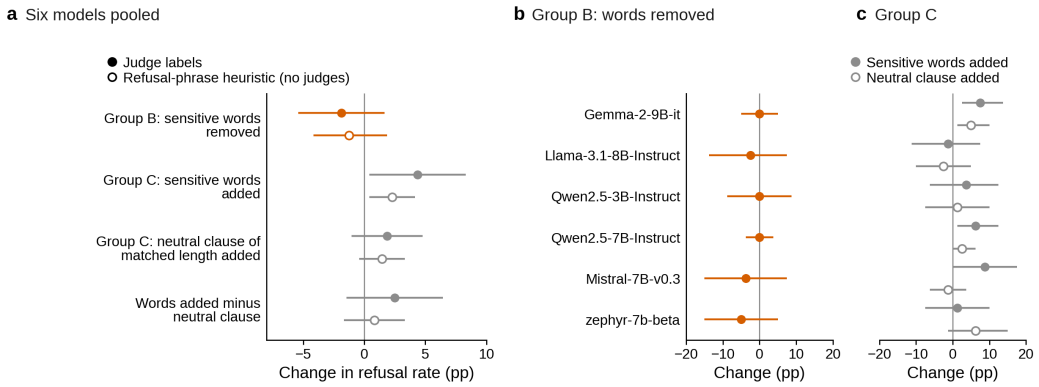


Figure 13. Lexical perturbation of the text prompts. (**a**) Change in refusal rate (percentage points) when the prompts are rewritten, pooled over the six models: judge-based estimate (filled) and refusal-phrase heuristic without the judges (open), with 95% cluster-bootstrap intervals. (**b**) Per-model change when the sensitive words are removed from benign prompts. (**c**) Per-model change for neutral prompts when sensitive words, or a neutral clause of matched length, are added.

Table 17. Lexical perturbation of the text prompts (six models pooled; 480 responses per arm). Each arm has one rewritten prompt per original item of its group. Refusal rate (%) of the original and the rewritten prompts (judge ensemble), and change in percentage points with 95% cluster-bootstrap intervals over the 80 items (bold: the interval excludes zero), once from the judge labels and once from the Korean refusal-phrase heuristic applied to the response text, which does not use the judges.

Group	Rewrite	Refusal Rate (%)		Change (pp) [95% CI]	
		Original	Rewritten	Judges	Phrases
B	Sensitive words removed	10.4	8.5	−1.9 [−5.4, 1.7]	−1.2 [−4.2, 1.9]
C	Sensitive words added	5.8	10.2	**+4.4 [0.4, 8.3]**	**+2.3 [0.4, 4.2]**
C	Neutral clause of matched length added	5.8	7.7	+1.9 [−1.0, 4.8]	+1.5 [−0.4, 3.3]
C	Words added minus neutral clause	–	–	+2.5 [−1.5, 6.5]	+0.8 [−1.7, 3.3]

Table 18. Per-model change in refusal rate (percentage points, 95% bootstrap intervals over 80 prompts; bold: the interval excludes zero) under the three lexical rewrites.

Model	Group B	Group C	
	Words Removed	Words Added	Neutral Clause Added
Gemma-2-9B-it	+0.0 [−5.0, 5.0]	**+7.5 [2.5, 13.8]**	**+5.0 [1.2, 10.0]**
Qwen2.5-7B-Instruct	+0.0 [−3.8, 3.8]	**+6.2 [1.2, 12.5]**	+2.5 [0.0, 6.2]
zephyr-7b-beta	−5.0 [−15.0, 5.0]	+1.2 [−7.5, 10.0]	+6.2 [−1.2, 15.0]
Llama-3.1-8B-Instruct	−2.5 [−13.8, 7.5]	−1.2 [−11.2, 7.5]	−2.5 [−10.0, 5.0]
Qwen2.5-3B-Instruct	+0.0 [−8.8, 8.8]	+3.8 [−6.2, 12.5]	+1.2 [−7.5, 10.0]
Mistral-7B-v0.3	−3.8 [−15.0, 7.5]	+8.8 [0.0, 17.5]	−1.2 [−6.2, 3.8]

5. Discussion

Across the text and image experiments, the tested models show a recurring pattern: uncertainty-related behavior looks more like all-or-none reactions to features of the input

than like graded, input-sensitive assessment. In the text domain, refusal rates do not move together across the harmful and benign groups, and refusals concentrate on some domains and not others, so that models are simultaneously too cautious on disguised-benign prompts and not cautious enough on some genuinely harmful ones, especially locale-specific harms. In the image domain, abstention is a comparatively well-behaved binary signal, firing much more often on out-of-domain inputs, but the stated confidence does not follow it: once a model decides to answer, the confidence it states is equivalent within about 3 to 5 points (90% intervals) whether the image was in- or out-of-distribution, although this comparison is conditional on answering and rests on 51 and 62 answered OOD trials. Diagnostic accuracy, stated confidence, and stated evidence location all behave poorly together on MedMNIST: near-chance performance paired with high, interpolation-insensitive, and in places inverted confidence, and evidence locations that agree with a CNN reference no more than position preferences alone would produce. In the text domain, a lexical perturbation did not support word-level triggers as the explanation (Section 4.12), so the refusals are better described as tied to the topic of a request than to its individual words; we did not paraphrase prompts or analyze internal representations, and the text and image experiments use different models and tasks, so the observations should be read as a consistent behavioral pattern and not as evidence for one latent deficit.

What the joint design adds is a common set of measures, not matched experiments. Table 19 lists what was applied in each modality. The text contrast matches prompts by domain and topic but not by wording, so intent and phrasing differ together (the lexical perturbation addresses the words of the prompt, not their syntax), whereas the vision contrast holds the whole prompt fixed and changes only the image; the references differ (an LLM-judge ensemble validated on one annotator's labels versus a trained CNN); and several controls exist in one modality only (the budget and wording controls and the lexical perturbation for text; the interpolation ablation and the scale sweep for vision). Within these limits two observations stand out that single-task studies would not show. First, in the vision experiment abstention and stated confidence dissociate: the VLMs detect gross domain mismatch but do not lower their confidence when they answer anyway. Second, the two low-cost interventions did not help in the settings tested: scale lowers stated confidence without improving discrimination (vision), and step-by-step prompting does not reliably help in either modality. The text and vision results on the second point are separate observations from different models and protocols, and their agreement does not indicate a shared mechanism.

Table 19. What was applied in each modality. The text and vision experiments share the measures listed in the first rows but differ in the contrast, the reference, and the controls, so they are separate observations and not matched experiments.

	Text (Six LLMs, 240 Korean Prompts)	**Vision (Two to Five VLMs, 205 Images)**
Contrast	Prompts matched by domain and topic; intent and wording differ	One fixed prompt and answer options; only the image differs
Reference	Two-judge LLM ensemble; 150 labels from one annotator	ConvMixer per task; Grad-CAM; no human labels
Step-by-step instruction	Submitted wording, two alternative wordings, a non-reasoning control, a token-budget control	One wording, one budget, no control
Surface-feature control	Lexical perturbation of the prompt words	Nearest-neighbor versus bicubic upscaling
Scale sweep	Not run (six models from different families, 3–9B)	Qwen3-VL 2B, 4B, 8B
Distribution shift	Not applicable	Gross OOD (satellite photographs); near-OOD not run

The comparison between VLMs and CNNs should be read with the difference in conditions in mind. The CNN is trained on each task's training set, whereas the VLMs are evaluated zero-shot with a single prompt, so the accuracy gap shows what an untrained general-purpose model can do under this protocol, not a ceiling for VLMs. Few-shot examples, task-specific prompting, or fine-tuning might improve VLM accuracy, and we did not test them. The CNN's role here is narrower: it shows that, for four of the six tasks, the 28×28 images contain enough information for a trained vision model (the ConvMixer is not above chance on the class-balanced samples of DermaMNIST and BreastMNIST), and it provides an independent saliency reference. Likewise, agreement with the CNN's Grad-CAM peak is a convergent-plausibility measure, not a test of the VLM's own faithfulness.

The scale sweep sharpens this picture without overturning it. Within the Qwen3-VL family, the overconfidence gap shrinks from 2B to 8B, but accuracy and the AUROC of confidence for correctness stay near chance, so the models become less confident without becoming better able to tell their correct answers from their incorrect ones. A practitioner hoping that a larger VLM within this range would resolve these issues should not expect it to: scale lowers the confidence numbers, not the underlying lack of diagnostic competence or discrimination.

The CoT results are practically important because CoT prompting is a common "free" intervention with a plausible mechanistic story (an explicit reasoning step should, in principle, let a model catch its own errors before committing to an answer). In our experiments this intuition did not hold on average: at a fixed generation budget, the instruction to assess intent and risk raised refusal of benign and neutral requests by 9.0 and 6.5 points across six LLMs and changed refusal of harmful requests by -2.3 points, generic step-by-step instructions changed nothing, and a response-lengthening instruction without reasoning lowered refusal of harmful requests by 8.1 points, so a lower refusal of harmful requests should not be read as an effect of reasoning; the effect of the risk-assessment wording is concentrated in some models (Section 4.11). For the smaller VLM, CoT left accuracy unchanged while increasing confidence, which is compatible with the reasoning text acting as a confidence-inflating artifact (a model that has just written several sentences of visual analysis, however ungrounded, may be more prone to expressing certainty) but does not demonstrate it.

For downstream evaluators, the practical implication is that a small LLM or VLM's stated confidence, stated reasoning, or refusal decision should not be used as a proxy for reliability without independent verification, and that neither of the two cheapest interventions (scale, CoT prompting) reliably closes the gap in the settings tested. More broadly, pairing a model under evaluation with a lightweight, cheaply trained task-specific reference model, as we do here (the judge ensemble for text, the ConvMixer CNN for images), is inexpensive to replicate for a specialized domain and surfaces the failure modes reported above without extensive human annotation, although the judge ensemble itself needs human validation: here it agreed with a blinded human labeler on the refusal decision but overstated benign-prompt refusal (Section 4.3).

6. Limitations

Judge validity. The text benchmark is labeled by a two-judge LLM ensemble. Inter-judge agreement is only fair to moderate (66.25% three-way, $\kappa = 0.376$; 75.07% binary, $\kappa = 0.479$), the two judges differ substantially in strictness, and about a quarter of the final labels were resolved by a keyword heuristic with limited recall. The qualitative findings are stable across aggregation rules (Section 4.2). Against 150 blinded human labels from a single annotator (the author), the primary labels reach $\kappa = 0.63$ [0.51, 0.73] on the refusal decision (Section 4.3); they over-call refusals, mainly at the boundary between compliance and

partial refusal (precision 67.7%), and overstate the pooled benign-prompt refusal rate by 8.0 pp [0.3, 16.2]. The sample holds 25 labeled responses per model, so the per-model rates in Table 1 are not validated individually; human–human agreement was not measured; and only the baseline responses were labeled, not the chain-of-thought responses, so the chain-of-thought changes in judge-labeled refusal have no human check. The model-level statements (the lowest group-A rate and the highest group-B rate of zephyr-7b-beta, five of six models on the Korea-specific domain) change with the labeling rule or are not resolved by resampling the prompts (Table 3), whereas the contrast between groups A and B holds under every rule; a second Korean-speaking annotator would be needed to measure human–human agreement. The Korean benchmark is hand-authored, and group-A items in different domains may differ in severity, which limits what the domain comparison can show about locale-specific under-protection.

Statistical precision. With 80 prompts per model and group and 205 images per VLM configuration, many single-model and subgroup estimates are imprecise. We report intervals, did not correct for multiple comparisons, and treat the p-values as descriptive.

OOD severity. Our OOD condition (satellite photographs presented as medical images) is an extreme, easily detectable domain mismatch, and the matched ID/OOD confidence comparison is restricted to the 51–62 answered OOD trials per model, which the model's own abstention selects; the stated confidence takes almost only the label "certain", which limits what an equivalence test on the encoded mean can show. We do not know whether the sharp abstention gap would persist for subtler shifts (an imaging modality or anatomical region absent from the model's training data, or realistic image corruptions), which is arguably the more practically important case.

Native resolution and task difficulty. The interpolation ablation compares two ways of upscaling the same 28 × 28 pixels and cannot show whether native resolution limits VLM accuracy; a control with higher-resolution medical images is needed. The VLMs are evaluated zero-shot with one prompt, and few-shot or fine-tuned settings were not tested.

Confidence encoding. We elicit confidence as one of five verbal labels and encode them numerically. The gap is positive under all five encodings tested, but its magnitude depends on the encoding, and for the 3B model, which used one label throughout, it is determined entirely by the encoding. Numeric or finer-grained elicitation was not tested.

CNN reference. One ConvMixer per task (seed 0, 10 epochs) serves as the reference. It is not above chance on the class-balanced samples of DermaMNIST and BreastMNIST (Section 4.6), so the VLM–CNN comparison is informative for four of the six tasks, and its Grad-CAM peaks refer to the CNN's predicted class, which is wrong for about a third of the images (the agreement result is unchanged on the images it classifies correctly).

Evidence-location measure. A 2 × 2 quadrant grid is coarse, and the measure compares the VLM's self-report with an independently trained CNN's saliency, not with the VLM's own internal computation, so it cannot show whether the VLM's stated evidence is faithful to its own decision process. For the Qwen3-VL 4B and 8B models, 55 and 50 of 205 responses lacked a parseable LOCATION field, concentrated in two tasks, so agreement for these models is bounded rather than estimated.

CoT design. For the text tasks we tested the submitted wording, two alternative reasoning formulations, and a non-reasoning control at a 1024-token budget, at which about a quarter of the benign and neutral responses (26% and 26%) still reached the limit; for the vision task we tested one wording at one budget, and the VLM result has no matched control. Few-shot CoT exemplars and self-consistency sampling [15] were not tested. The new text conditions are labeled by the judge ensemble only, which overstates benign-prompt refusal against the human labels (Section 4.3), and they were generated on a different GPU than

the submitted run (the refusal decision of the final label agreed for 89.6% of re-run baseline responses), so the conclusions apply to this protocol and to judge-based labels.

Mechanism. The lexical perturbation changed word choice in 80 prompts per arm, so its pooled intervals are about 7 points wide; the rewrites were written by the author and keep the topic of each request, and no group-A prompt was rewritten. Paraphrases that change the topic, representation-level analyses, and the VLM confidence behavior were not examined, so the interpretation of the VLM results in terms of surface features remains a description of behavior.

Model and task coverage. We test five VLM configurations (2–8B) from two model families (Qwen2.5-VL, Qwen3-VL) and six LLMs (3–9B) on six MedMNIST tasks and 240 Korean prompts. Larger scales, other model families, closed-weight models, and full-resolution clinical imagery remain untested, and the text and image experiments use different models, contrasts, references and controls (Table 19).

7. Conclusions

We presented a controlled audit, with matched comparisons in each modality, of uncertainty-expression failures spanning Korean-language text safety, medical-image out-of-distribution detection, VLM confidence calibration, and explanation grounding in small open-weight models, together with two candidate fixes. Across the experiments the same qualitative pattern recurs: behavior that looks reasonable in aggregate (models refuse most genuinely harmful text; VLMs abstain much more on OOD images; larger models state lower confidence) but that, on closer inspection, behaves, in our tests, more like all-or-none reactions to the topic of a request or the domain of an image than like input-appropriate graded judgment, and that fails most visibly on locale-specific harms, answered OOD inputs, and low-resolution specialized imagery. Neither moving to a larger model within one family nor a single step-by-step prompt reliably closes the gap: scale lowered stated confidence without improving accuracy or discrimination, and the step-by-step prompt raised refusal of benign and neutral requests across the six LLMs only when it asked the model to assess intent and risk (generic step-by-step prompts did not) and increased overconfidence in the smaller VLM. We hope this framing motivates evaluation practices that test both directions of the refusal boundary, pair model self-reports with an independently trained reference wherever possible, validate LLM judges against human annotation, and treat "use a bigger model" or "add a reasoning step" as empirical claims to be tested rather than assumed fixes.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/electronics15215158/s1, File S1 (`text_benchmark/`): the 240 Korean prompts, the judge rubric, the raw responses of the six LLMs with and without the step-by-step instruction, the labels of both judges, the final labels and the refusal-rate tables; File S2 (`vlm/`): the manifests of the 205 sampled MedMNIST images and of the in-distribution and out-of-distribution sets, the raw VLM outputs, the Grad-CAM peak quadrants, the ConvMixer reference results and the summary tables; File S3 (`text_benchmark_controls/`): the budget-matched controls and the lexical perturbation of the text prompts (rewritten prompts, merged judge labels and raw judge outputs); File S4 (`human_validation/`): the blinded human labels, the merged human and judge labels and the weighted agreement estimates; File S5 (`analysis_outputs/` and `code/`): the outputs of the statistical analyses and the scripts that produced the data, analyses and figures.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study did not involve human or animal subjects; all image data used are publicly available, de-identified benchmark datasets (MedMNIST v2, RESISC45). The only human annotation, a blinded labeling of 150 model-generated responses, was performed by the author; no participants were recruited and no personal data were collected.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Korean 240-item over-refusal prompt set with the judging rubric, all model responses, judge outputs and final labels, the 150 blinded human labels, the control and perturbation prompts with their responses and labels, VLM outputs, Grad-CAM quadrant files, the statistical analysis outputs, and the analysis code used in this study are provided as Supplementary Materials. MedMNIST v2 [19] and RESISC45 [20] are publicly available benchmark datasets and are not redistributed.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback. *arXiv* **2022**, arXiv:2203.02155. [CrossRef]
2. Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv* **2022**, arXiv:2204.05862. [CrossRef]
3. Röttger, P.; Kirk, H.R.; Vidgen, B.; Attanasio, G.; Bianchi, F.; Hovy, D. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 5377–5400. [CrossRef]
4. Cui, J.; Chiang, W.-L.; Stoica, I.; Hsieh, C.-J. OR-Bench: An Over-Refusal Benchmark for Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*; JMLR: Norfolk, MA, USA, 2025; Volume 267, pp. 11515–11542. Available online: https://proceedings.mlr.press/v267/cui25a.html (accessed on 6 October 2026).
5. Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *arXiv* **2023**, arXiv:2306.00890. [CrossRef]
6. Moor, M.; Banerjee, O.; Abad, Z.S.H.; Krumholz, H.M.; Leskovec, J.; Topol, E.J.; Rajpurkar, P. Foundation Models for Generalist Medical Artificial Intelligence. *Nature* **2023**, *616*, 259–265. [CrossRef] [PubMed]
7. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*; JMLR: Norfolk, MA, USA, 2017; Volume 70, pp. 1321–1330. Available online: https://proceedings.mlr.press/v70/guo17a.html (accessed on 6 October 2026).
8. Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. Language Models (Mostly) Know What They Know. *arXiv* **2022**, arXiv:2207.05221. [CrossRef]
9. Jacovi, A.; Goldberg, Y. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 4198–4205. [CrossRef]
10. Deng, Y.; Zhang, W.; Pan, S.J.; Bing, L. Multilingual Jailbreak Challenges in Large Language Models. *arXiv* **2023**, arXiv:2310.06474. [CrossRef]
11. Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, X.; Wen, J.-R. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, US, 2023; pp. 292–305. [CrossRef]
12. Cheng, L.; Zhang, H.; Di, B.; Niyato, D.; Song, L. Large Language Models Empower Multimodal Integrated Sensing and Communication. *IEEE Commun. Mag.* **2025**, *63*, 190–197. [CrossRef]
13. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity Checks for Saliency Maps. *arXiv* **2018**, arXiv:1810.03292. [CrossRef]
14. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 618–626. [CrossRef]
15. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022. Available online: https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html (accessed on 6 October 2026).
16. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 121–127. [CrossRef] [PubMed]
17. Mun, C.; Ha, H.; Lee, O.; Cheon, M. Enhancing AI-CDSS with U-AnoGAN: Tackling Data Imbalance. *Comput. Methods Programs Biomed.* **2024**, *244*, 107954. [CrossRef] [PubMed]

18. Zhang, M.; Lu, J.; Cheng, L.; Yang, X.; Zhou, J.; Wan, M. Enhancing Infant Cry Recognition Using Lightweight CNN with Hybrid Feature Augmentation. *Biomed. Signal Process. Control* **2026**, *121*, 110367. [CrossRef]
19. Yang, J.; Shi, R.; Wei, D.; Liu, Z.; Zhao, L.; Ke, B.; Pfister, H.; Ni, B. MedMNIST v2—A Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification. *Sci. Data* **2023**, *10*, 41. [CrossRef] [PubMed]
20. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
21. Trockman, A.; Kolter, J.Z. Patches Are All You Need? *Trans. Mach. Learn. Res.* **2023**. Available online: https://openreview.net/forum?id=rAnB7JSMXL (accessed on 6 October 2026). [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.